Cross-Lingual Transfer with Target Language-Ready Task Adapters

Marinela Parović[1] **Alan Ansell**[1] **Ivan Vulić**[1] **Anna Korhonen**[1]
[1]Language Technology Lab, TAL, University of Cambridge
{mp939,aja63,iv250,alk23}@cam.ac.uk

Abstract

Adapters have emerged as a modular and parameter-efficient approach to (zero-shot) cross-lingual transfer. The established MAD-X framework employs separate language and task adapters which can be arbitrarily combined to perform the transfer of any task to any target language. Subsequently, BAD-X, an extension of the MAD-X framework, achieves improved transfer at the cost of MAD-X's modularity by creating 'bilingual' adapters specific to the source-target language pair. In this work, we aim to take the best of both worlds by (i) fine-tuning *task* adapters adapted to the target language(s) (so-called *'target language-ready' (TLR)* adapters) to maintain high transfer performance, but (ii) without sacrificing the highly modular design of MAD-X. The main idea of 'target language-ready' adapters is to resolve the training-vs-inference discrepancy of MAD-X: the task adapter 'sees' the target language adapter for the very first time during inference, and thus might not be fully compatible with it. We address this mismatch by exposing the task adapter to the target language adapter during training, and empirically validate several variants of the idea: in the simplest form, we alternate between using the source and target language adapters during task adapter training, which can be generalized to cycling over any set of language adapters. We evaluate different TLR-based transfer configurations with varying degrees of generality across a suite of standard cross-lingual benchmarks, and find that the most general (and thus most modular) configuration consistently outperforms MAD-X and BAD-X on most tasks and languages.

1 Introduction and Motivation

Recent progress in multilingual NLP has mainly been driven by massively multilingual Transformer models (MMTs) such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and mT5 (Xue et al., 2021), which have been trained on the unlabeled data of 100+ languages. Their shared multilingual representation spaces enable zero-shot cross-lingual transfer (Pires et al., 2019; K et al., 2020), that is, performing tasks with a reasonable degree of accuracy in languages that entirely lack training data for those tasks.

Zero-shot cross-lingual transfer is typically performed by fine-tuning the pretrained MMT on task-specific data in a high-resource *source* language (i.e., typically English), and then applying it directly to make task predictions in the *target* language. In the standard setup, the model's knowledge about the target language is acquired solely during the pretraining stage (Artetxe et al., 2020). In order to improve the transfer performance, task fine-tuning can be preceded with fine-tuning on unlabeled data in the target language (Ponti et al., 2020; Pfeiffer et al., 2020b). Nonetheless, the performance on the target languages in such scenarios is lower than that on the source language, and the difference is known as the *cross-lingual transfer gap* (Hu et al., 2020). Crucially, the transfer gap tends to increase for the languages where such transfer is needed the most (Joshi et al., 2020): i.e., for low-resource target languages, and languages typologically more distant from the source language (e.g., English) (Lauscher et al., 2020).

Adapters (Rebuffi et al., 2017; Houlsby et al., 2019) have emerged as a prominent approach for aiding zero-shot cross-lingual transfer (Pfeiffer et al., 2020b; Üstün et al., 2022a; Ansell et al., 2021; Parović et al., 2022). They offer several benefits: (i) providing additional representation capacity for target languages; (ii) much more parameter-efficient fine-tuning compared to full-model fine-tuning, as they allow the large MMT's parameters to remain unmodified, and thus preserve the multilingual knowledge the MMT has acquired during pretraining. They also (iii) provide modularity in learning and storing different facets of knowledge (Pfeiffer et al., 2020a): this property enables them to be combined in favorable ways to achieve better

Findings of the Association for Computational Linguistics: ACL 2023, pages 176–193
July 9-14, 2023

performance, and previously fine-tuned modules (e.g., language adapters) to be reused across different applications.

The established adapter-based cross-lingual transfer framework MAD-X (Pfeiffer et al., 2020b) trains separate language adapters (LAs) and task adapters (TAs) which can then be arbitrarily combined for the transfer of any task to any language. Despite having a highly modular design, stemming primarily from dedicated per-language and per-task adapters, MAD-X's TAs lack 'adaptivity' to the target language(s) of interest: i.e., its TAs are fully *target language-agnostic*. More precisely, during task fine-tuning, the MAD-X TA is exposed only to the source language LA, and 'sees' the target language TA and examples from that language for the first time only at inference. This deficiency might result in incompatibility between the TA and the target LA, which would emerge only at inference.

BAD-X (Parović et al., 2022) trades off MAD-X's high degree of modularity by introducing *'bilingual'* language adapters specialized for transfer between the source-target language pair.[1] While such transfer direction specialization results in a better performance, the decrease in modularity results in much larger computational requirements: BAD-X requires fine-tuning a dedicated bilingual LA for every language pair of interest followed up by fine-tuning a dedicated TA again for each pair.

Prior work has not explored whether this specialization (i.e., exposing the target language at training time) can be done successfully solely at the level of TAs whilst preserving modularity at the LA level. Such specialization in the most straightforward bilingual setup still requires fine-tuning a dedicated TA for each target language of interest. However, this is already a more pragmatic setup than BAD-X since TAs are much less computationally expensive to train than LAs. Moreover, as we show in this work, it is possible to also extend TA fine-tuning to more target languages, moving from bilingual specialization to the more universal multilingual 'exposure' and towards *multilingual language-universal TAs*.

In this work, we aim to create a modular design inspired by MAD-X while seeking to reap the benefits of the exposure to one or more target languages. To this end, we thus introduce *target language-ready (TLR)* task adapters designed to excel at a particular target language or at a larger set of target languages. In the simplest bilingual variant, TLR TAs are trained by alternating between source and target LAs, while the more general version allows cycling over any set of LAs. Creating TLR TAs does not require any expensive retraining or alternative training of LAs.

We run experiments with a plethora of standard benchmarks focused on zero-shot cross-lingual transfer and low-resource languages, covering 1) NER on MasakhaNER; 2) dependency parsing (DP) on Universal Dependencies; 3) natural language inference (NLI) on AmericasNLI and XNLI; 4) QA on XQuAD and TyDiQA-GoldP. Our results show that TLR TAs outperform MAD-X and BAD-X on all tasks on average, and offer consistent gains across a large majority of the individual target languages. Importantly, the most general TLR TA, which is shared between all target languages and thus positively impacts modularity and reusability, shows the strongest performance across the majority of tasks and target languages. Fine-tuning the TA in such multilingual setups also acts as a *multilingual regularization* (Ansell et al., 2021): while the TA gets exposed to different target languages (i.e., maintaining its TLR property), at the same time it does not overfit to a single target language as it is forced to adapt to more languages, and thus learns more universal cross-language features. Our code and models are publicly available at: `https://github.com/parovicm/tlr-adapters`.

2 Methodology

2.1 Background

Adapters. Following MAD-X and BAD-X, in this work we focus on the most common adapter architecture, *serial adapters* (Houlsby et al., 2019; Pfeiffer et al., 2021a), but we remind the reader that other adapter options are available (He et al., 2022) and might be used in the context of cross-lingual transfer. Serial adapters are lightweight bottleneck modules inserted within each Transformer layer. The architecture of an adapter at each layer consists of a down-projection, a non-linearity and an up-projection followed by a residual connection. Let the down-projection at layer l be a matrix $\mathbf{D}_l \in \mathbb{R}^{h \times d}$ and the up-projection be a matrix $\mathbf{U}_l \in \mathbb{R}^{d \times h}$ where h is the hidden size of the Transformer and d is the hidden size of the adapter. If we denote the hidden state and the residual at layer l as $\mathbf{h}_l$ and $\mathbf{r}_l$ respectively, the adapter computation

[1] Similarly, such bilingual adapters have been used in multilingual NMT research to boost translation between particular language pairs (Bapna and Firat, 2019; Philip et al., 2020).

of layer l is then given by:

$$A_l(\mathbf{h}_l, \mathbf{r}_l) = \mathbf{U}_l(\text{ReLU}(\mathbf{D}_l(\mathbf{h}_l))) + \mathbf{r}_l, \quad (1)$$

with ReLU as the activation function.

MAD-X and BAD-X Frameworks. MAD-X trains dedicated LAs and TAs (Pfeiffer et al., 2020b). LAs are trained using unlabeled Wikipedia data with a masked language modeling (MLM) objective. TAs are trained using task-specific data in the source language. Given a source language L_s and a target language L_t, MAD-X trains LAs for both L_s and L_t. The TA is trained while stacked on top of the L_s LA, which is frozen. To make predictions on L_t, the L_s LA is swapped with the L_t LA.

Unlike MAD-X, which is based on monolingual adapters, BAD-X trains bilingual LAs (Parović et al., 2022). A bilingual LA is trained on the unlabeled data of both L_s and L_t and the TA is then trained on task-specific data in L_s, stacked on top of the bilingual LA. To perform inference on the task in L_t, the same configuration is kept since the bilingual LA 'knows' both L_s and L_t.

2.2 Target Language-Ready Task Adapters

Instead of sacrificing the LAs' modularity as in BAD-X, it might be more effective to keep MAD-X's language-specific LAs and opt to prepare only the TAs to excel at a particular target language L_t, or a set of target languages of interest. Assuming LAs are available for the source language L_s and K target languages $L_{t,i},\ i = 1, \ldots, K$, we cycle over all $K + 1$ LAs during TA training, resulting in the so-called *multilingual TLR TA*. This general idea is illustrated in Figure 1. The bilingual variant with a TLR TA trained by alternating between the source and target LA is a special case of the multilingual variant where $K = 1$, while the original MAD-X setup is obtained by setting $K = 0$.[2]

This procedure exposes a single target language (bilingual TLR TA) or multiple target languages (multilingual TLR TA) to the TA as soon as its fine-tuning phase, making it better equipped (i.e., *ready*) for the inference phase, where the TA is combined with the single L_t LA.

TLR Variants. While BILINGUAL TA fine-tuning follows naturally from BAD-X, and it seems suitable for transfer between a fixed pair of L_s and L_t, it might be better to train the TA only on top of the L_t LA. Such TARGET-only TLR TAs could be particularly effective for higher-resource languages whose LAs have been trained on sufficient corpora, to the extent that pairing them with L_s is detrimental. This could be especially detectable for higher-resource L_t-s that are also distant from L_s or lack adequate vocabulary overlap with it.

TARGET and BILINGUAL TLR TAs require training of dedicated TAs for every L_t of interest, which makes them computationally less efficient than MAD-X, and they introduce more parameters overall. Using MULTILINGUAL TLR TAs mitigates this overhead. We consider two variants of MULTILINGUAL TAs. First, the so-called TASK-MULTI TLR variant operates over the source language and the set of all target languages available for

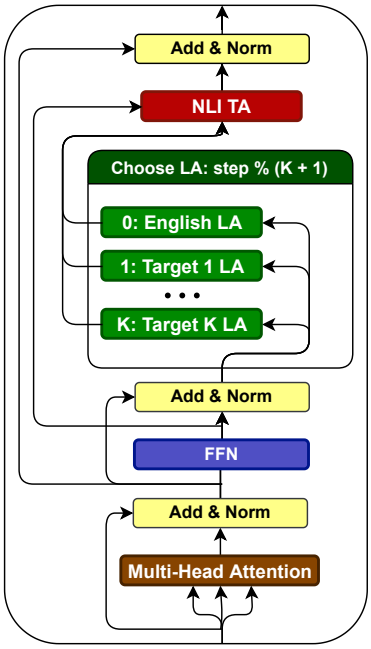

Figure 1: A general *multilingual* task adapter (TA) *target language-ready* (TLR) module at one MMT layer, showing the language adapters (LAs) for English as the source language and K target languages along with the NLI TA. The TA is trained by cycling over the $K + 1$ LAs associated with the $K + 1$ languages. For a given *step* number, only the LA *step* $\%$ $(K + 1)$ is switched on and the forward pass goes through that LA. Setting $K = 0$ results in the original MAD-X setup, where only the source LA is switched on, while a bilingual TLR variant is given by $K = 1$. Setting $K = 1$ and removing the English LA formulates the TARGET-only TLR variant. See §2.2 for the descriptions of all the variants. The same adapter configuration(s), but with different parameters, are added at each MMT layer.

[2]It is also possible to train a TA directly without relying on any LA at all. However, previous research (Ansell et al., 2021) has empirically validated that this *'TA-only'* variant is consistently outperformed by MAD-X; hence, we do not discuss nor compare to 'TA-only' in this work.

the task under consideration (e.g., all languages represented in the MasakhaNER dataset). Second, the ALL-MULTI TLR variant combines the source language with all target languages across datasets of multiple tasks (e.g., all languages represented in MasakhaNER, all languages represented in AmericasNLI, etc.); see §3 later. These variants increase modularity and parameter efficiency and are as modular and parameter-efficient as MAD-X per each task: a single TA is required to handle transfer to any target language. At the same time, unlike MAD-X, they are offered some exposure to the representations arising from the multiple target languages they will be used for. Handling multiple LAs at fine-tuning might make the TAs more robust overall: multilinguality might act as a regularization forcing the TA to focus on more universal cross-language features (Ansell et al., 2021).

3 Experimental Setup

Evaluation Tasks and Languages. We comprehensively evaluate our TLR adapter framework on a suite of standard cross-lingual transfer benchmarks. They span four different task families (NER, DP, NLI and QA), with a total of six different datasets and 35 different target languages, covering a typologically and geographically diverse language sample of both low- and high-resource languages.

For NER, we use the MasakhaNER dataset (Adelani et al., 2021) which contains 10 low-resource languages from the African continent.[3] For DP, we use Universal Dependencies 2.7 (Zeman et al., 2020) and inherit the set of 10 typologically diverse low-resource target languages from BAD-X (Parović et al., 2022). For NLI, we rely on the AmericasNLI dataset (Ebrahimi et al., 2022), containing 10 low-resource languages from the Americas, as well as a subset of languages from XNLI (Conneau et al., 2018). Finally, for QA we use subsets of languages from XQuAD (Artetxe et al., 2020) and TyDiQA-GoldP (Clark et al., 2020). The subsets for XNLI, XQuAD and TyDiQA-GoldP were selected to combine (i) low-resource languages (Joshi et al., 2020), with (ii) higher-resource languages for which dedicated (i.e., 'MAD-X') LAs were readily available. The full overview of all tasks, datasets, and languages with their language codes is provided in Table 5 in Appendix A.

[3]We exclude Amharic from our experiments as it uses a script not supported by mBERT, resulting in 9 NER target languages.

	NER	DP	NLI	QA
Batch Size	8	8	32	16
Epochs	10	10	5	15
Learning Rate	$5 \cdot 10^{-5}$	$5 \cdot 10^{-5}$	$2 \cdot 10^{-5}$	10^{-4}
Eval Freq. (steps)	250	250	625	625
Eval Metric	F1	LAS	Acc	F1

Table 1: Hyperparameters for different tasks.

Underlying MMT. We report results on all tasks with mBERT, pretrained on Wikipedias of 104 languages (Devlin et al., 2019). mBERT has been suggested by prior work as a better-performing MMT for truly low-resource languages (Pfeiffer et al., 2021b; Ansell et al., 2021). To validate the robustness of our TLR adapters, we also use XLM-R (Conneau et al., 2020) for a subset of tasks.

Language Adapters. We train LAs for the minimum of 100 epochs or 100,000 steps with a batch size of 8, a learning rate of $5 \cdot 10^{-5}$ and a maximum sequence length of 256.[4] We evaluate the LAs every 1,000 steps for low-resource languages and every 5,000 steps for high-resource ones, and choose the LA that yields the lowest perplexity, evaluated on the 5% of the held-out monolingual data (1% for high-resource languages). For the BAD-X baseline, we directly use the bilingual LAs from (Parović et al., 2022). Following Pfeiffer et al. (2020b), the adapter reduction factor (i.e., the ratio between MMT's hidden size and the adapter's bottleneck size) is 2 for all LAs. For the MAD-X LAs, we use the efficient Pfeiffer adapter configuration (Pfeiffer et al., 2020a) with invertible adapters, whereas BAD-X LAs do not include them.

Task Adapters. We fine-tune TAs by stacking them on top of the corresponding LAs (see Figure 1). During their fine-tuning, the MMT's parameters and all the LAs' parameters are frozen. The adapter reduction factor for all TAs is 16 as in prior work (Pfeiffer et al., 2020b) (i.e., $d = 48$), and, like the LAs, they use the Pfeiffer configuration. The hyperparameters across different tasks, also borrowed from prior work, are listed in Table 1. In addition, we use early stopping of 4 when training the QA TA (i.e., we stop training when the F1 score does not increase for the four consecutive evaluation cycles). We use the English SQuADv1.1 training data (Rajpurkar et al., 2016) for TyDiQA-GoldP since (i) it is much larger than TyDiQA's

[4]For some low-resource languages with small corpora 100 epochs leads to under-training, so the minimum number of training steps is set to 30,000.

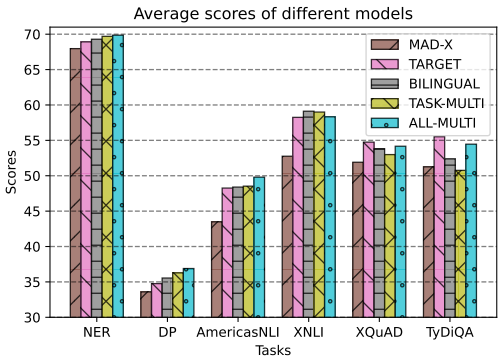

Figure 2: The average scores of MAD-X, TARGET, BILINGUAL, TASK-MULTI and ALL-MULTI variants on NER (F1), DP (LAS), AmericasNLI (acc), XNLI (acc), XQuAD (F1) and TyDiQA (F1) datasets.

native training set, and (ii) we observed higher performance on target languages in our preliminary experiments than with TyDiQA's training data.

Transfer Setup: Details. In all our transfer experiments, the source language L_s is fixed to English, and we evaluate different variants described in §2.2. For the MAD-X baseline, we rely on its 'MAD-X v2.0' variant, which drops the adapters in the last layer of the Transformer, which has been found to improve transfer performance across the board (Pfeiffer et al., 2021b). For the TASK-MULTI TLR variant, along with using the English LA, we fine-tune TAs using the LAs of all our evaluation languages in that particular dataset. For instance, for DP this spans 10 languages, while for NLI, we fine-tune a separate TASK-MULTI TLR with the 10 languages from AmericasNLI, and another one for the XNLI languages. For the ALL-MULTI TLR variant, in addition to English LA, we cycle over the LAs of all our evaluation languages from all the tasks and datasets.

4 Results and Discussion

Main Results. The main results with mBERT for all tasks and all languages are shown in Table 2, with the averages concisely provided in Figure 2. Additional results with XLM-R are available in Appendix B. As a general trend, we observe that all proposed TLR variants outperform MAD-X on the majority of the target languages across all tasks. Besides reaching higher averages on all tasks, the best per-task variants from the TLR framework surpass MAD-X on: 9/9 (NER), 10/10 (DP), 10/10 (AmericasNLI), 6/6 (XNLI), 4/4 (XQuAD) and 5/5 (TyDiQA) target languages. We also demonstrate that gains are achieved over the much less modular BAD-X on two tasks (DP, AmericasNLI) for which we had readily available BAD-X LAs. In sum, the comprehensive set of results from Table 2 confirms the effectiveness and versatility of TLR adapters across a range of (typologically diverse) target languages and datasets.

Breakdown of Results across Tasks and TLR Variants. On NER and DP we observe very similar trends in results. Importantly, the most modular ALL-MULTI variant offers the highest performance overall: e.g., it reaches the average F1 score of 69.86% in the NER task, while outperforming MAD-X by 1.9% on average and on all 9 target languages. Pronounced gains with that variant are also indicated in the DP task. The TARGET and BILINGUAL variants also yield gains across the majority of languages, with BILINGUAL being the stronger of the two. However, their overall utility in comparison to ALL-MULTI is lower, given their lower performance coupled with lower modularity.

On AmericasNLI, all TLR variants display considerable gains over MAD-X, achieving 5-6% higher average accuracy. They outperform MAD-X on all 10 target languages, except the TASK-MULTI variant with only a slight drop on AYM. The best variant is once again the most modular ALL-MULTI variant, which is better than the baselines and all the other variants on 6/10 target languages.

On XNLI, which involves some higher-resource languages such as AR, HI and ZH, all TLR variants reach higher average accuracy than MAD-X. The gains peak around 5-6% on average; however, this is due mainly to SW where MAD-X completely fails, achieving the accuracy of random choice. Nonetheless, the TLR variants attain better scores on all other languages as well (the only exception is ALL-MULTI on AR). Besides SW, TH also marks a large boost of up to 11.2% with the BILINGUAL variant, while the other languages attain more modest gains of up to 2%. We remark that the BILINGUAL variant now obtains the highest average accuracy: we speculate that this could be a consequence of target languages now being on the higher-resource end compared to MasakhaNER and AmericasNLI.

Our final task family, QA, proves yet again the benefits of transfer with TLR adapters. On XQuAD and TyDiQA-GoldP, the best TLR variant is now the TARGET adapter. This might be partially due to a good representation of high-resource languages

Method	HAU	IBO	KIN	LUG	LUO	PCM	SWA	WOL	YOR	**avg**	Better
MAD-X	81.30	70.27	62.53	64.70	48.20	72.94	74.20	65.56	71.95	67.96	
TARGET	77.58	**73.99**	64.34	68.08	51.20	74.00	75.26	63.04	72.76	68.92	7/9
BILINGUAL	79.93	71.90	64.74	**68.68**	51.18	74.82	75.68	63.68	**73.00**	69.29	7/9
TASK-MULTI	81.83	72.76	65.03	66.95	50.69	75.35	**76.59**	65.87	72.26	69.70	9/9
ALL-MULTI	**82.39**	71.82	**65.12**	66.38	**51.38**	**76.17**	76.42	**66.93**	72.10	**69.86**	9/9
LEAVE-OUT-TASK	82.54	70.88	65.74	65.78	49.93	75.33	76.10	65.27	72.61	69.35	8/9
LEAVE-OUT-TARG	82.60	71.11	64.50	66.95	51.38	75.21	75.62	65.57	71.90	69.43	8/9

(a) NER: F1

Method	AF	BM	EU	KPV	MR	MT	MYV	TE	UG	WO	**avg**	Better
MAD-X	55.21	13.73	33.20	23.12	26.18	47.42	35.70	49.62	19.60	32.07	33.59	
BAD-X	54.54	11.92	31.45	22.55	26.56	43.52	39.31	46.22	15.24	35.28	32.66	
TARGET	56.91	13.62	34.55	21.96	28.05	45.63	38.47	51.80	17.22	39.41	34.76	6/10
BILINGUAL	56.86	14.25	33.56	22.84	27.71	48.46	38.67	53.56	19.74	39.82	35.55	9/10
TASK-MULTI	56.56	15.43	34.90	22.93	**28.70**	51.85	39.18	53.51	19.48	40.29	36.28	8/10
ALL-MULTI	**57.11**	**15.46**	**35.32**	**23.76**	28.35	**53.68**	**39.71**	**53.83**	**20.32**	**41.34**	**36.89**	10/10
LEAVE-OUT-TASK	56.99	16.40	33.88	25.27	28.28	55.03	39.96	54.11	21.52	40.41	37.19	10/10
LEAVE-OUT-TARG	56.97	15.87	35.67	25.47	27.82	53.93	39.68	52.54	20.95	40.65	36.95	10/10

(b) DP: LAS

Method	AYM	BZD	CNI	GN	HCH	NAH	OTO	QUY	SHP	TAR	**avg**	Better
MAD-X	50.40	40.93	37.47	55.60	38.27	46.61	39.71	48.80	38.27	38.80	43.49	
BAD-X	46.13	44.67	45.87	56.80	44.93	47.70	41.71	47.87	**49.07**	39.47	46.42	
TARGET	50.53	**47.20**	44.13	58.00	43.73	**50.54**	41.04	55.87	46.13	45.47	48.26	10/10
BILINGUAL	**51.73**	46.80	43.07	58.53	**46.13**	48.51	43.32	55.47	46.00	44.40	48.40	10/10
TASK-MULTI	49.60	45.60	44.67	58.67	46.00	50.27	43.32	55.87	47.07	44.27	48.53	9/10
ALL-MULTI	51.33	**47.20**	**47.20**	**60.00**	46.00	48.10	**45.59**	**58.40**	48.00	**46.13**	**49.80**	10/10
LEAVE-OUT-TASK	54.40	42.80	44.40	58.13	42.40	47.56	41.44	56.80	42.80	43.73	47.45	10/10
LEAVE-OUT-TARG	51.07	44.27	47.33	59.47	44.53	47.43	43.98	56.53	46.53	42.93	48.41	10/10

(c) AmericasNLI: accuracy

Method	AR	HI	SW	TH	UR	ZH	**avg**	Better
MAD-X	62.75	56.75	33.33	43.75	56.41	63.57	52.76	
TARGET	62.87	57.92	53.93	52.08	56.79	**65.93**	58.25	6/6
BILINGUAL	63.49	**58.62**	54.71	**54.95**	**57.47**	65.49	**59.12**	6/6
TASK-MULTI	**64.07***	57.88	**55.35**	54.19	56.81	65.69	59.00	6/6
ALL-MULTI	61.98	57.80	54.15	53.25	57.05	65.75	58.33	5/6

(d) XNLI: accuracy

Method	AR	HI	TH	ZH	**avg**	Better
MAD-X	58.97/42.27	51.09/36.47	40.45/30.59	57.12/46.72	51.91/39.01	
TARGET	60.40/43.95	**54.91**/40.59	**44.95/36.22**	58.73/48.24	**54.75/42.25**	4/4
BILINGUAL	**60.44/44.29**	54.18/40.42	42.68/33.95	57.95/48.32	53.81/41.75	4/4
TASK-MULTI	59.04/43.28	52.03/37.56	41.91/31.43	**58.97**/48.91	52.99/40.30	4/4
ALL-MULTI	58.67/42.44	54.79/**41.42**	44.67/35.97	58.57/**48.99**	54.17/42.20	3/4

(e) XQuAD: F1/EM

Method	AR	BN	SW	TE	TH	**avg**	Better
MAD-X	51.10/34.42	56.21/42.48	55.04/42.49	46.56/34.53	47.41/32.91	51.26/37.37	
TARGET	**56.88/40.93**	**59.47/49.56**	**61.91/50.10**	**49.92**/39.31	49.36/34.81	**55.51/42.94**	5/5
BILINGUAL	53.50/38.65	53.47/40.71	58.26/49.10	48.47/38.12	48.22/33.67	52.38/40.05	4/5
TASK-MULTI	49.33/34.42	50.92/39.82	58.34/48.70	49.30/**39.76**	45.93/33.67	50.76/39.27	2/5
ALL-MULTI	55.26/39.41	55.17/41.59	60.42/49.30	49.35/38.86	**52.09/39.62**	54.46/41.76	4/5

(f) TyDiQA: F1/EM

Table 2: Results of all methods and TLR variants on all tasks and target languages. The highest task score per each language in **bold**, but excluding the two ablation subvariants of ALL-MULTI placed below the dashed horizontal lines (LEAVE-OUT-TASK and LEAVE-OUT-TARG). *Better* refers to the number of target languages for which each TLR variant scores higher than MAD-X. An asterisk (*) next to the best TLR variant indicates *non-significant* gains over MAD-X, where the significance analysis has been conducted using Student's t-test with $p = 0.05$.

Method	DP	AmericasNLI
MAD-X	31.29	45.33
BAD-X	32.66	46.42
TARGET	35.15	48.24
BILINGUAL	34.41	48.47
TASK-MULTI	35.86	48.05
ALL-MULTI	36.47	48.49

Table 3: Robustness of TLR adapters. Average scores on DP and AmericasNLI when MAD-X LAs are trained with a different configuration and training setup. Per-language scores are available in Appendix C.

Method	NER	AmericasNLI
MAD-X	68.27	44.66
TARGET	68.49	47.92
BILINGUAL	69.24	48.32
TASK-MULTI	69.47	48.55
ALL-MULTI	69.10	49.10
LEAVE-OUT-TASK	69.37	47.96
LEAVE-OUT-TARG	69.13	48.44

Table 4: Gains with TLR adapters over MAD-X persist when scores are averages across 3 runs (i.e. 3 different random seeds). Average scores reported, while per-language scores are provided in Appendix D.

such as AR, HI, or ZH in mBERT and its subword vocabulary. However, we observe gains with TARGET also on lower-resource languages such as BN and SW on TyDiQA, which might indicate that the higher complexity of the QA task is at play in comparison to tasks such as NER and NLI.

Crucially, the most modular ALL-MULTI TLR variant, which trains a single TA per each task, yields very robust and strong performance across all tasks (including the two QA tasks) and both on high-resource and low-resource languages.

Towards Language-Universal Task Adapters? Strictly speaking, if a new $(K+1)$-th target language is introduced to our proposed TLR framework, it would be necessary to train the multilingual TLR TA anew to expose it to the new target language. In practice, massively multilingual TAs could still be applied even to languages 'unseen' during TA fine-tuning (e.g., in the same way as the original MAD-X framework does). This violates the TLR assumption, as the TA sees the target language only at inference. However, this setup might empirically validate another desirable property of our multilingual TLR framework from Figure 1: exposing the TA at fine-tuning to a multitude of languages (and their corresponding LAs) might equip the TA with improved transfer capability even to unseen languages. Put simply, the TA will not overfit to a single target language or a small set of languages as it must learn to balance across a large and diverse set of languages; see §2.

We thus run experiments on MasakhaNER, UD DP, and AmericasNLI with two subvariants of the most general ALL-MULTI variant. First, in the LEAVE-OUT-TASK subvariant, we *leave out* all the LAs for the languages from the corresponding task dataset when fine-tuning the TA: e.g., for AmericasNLI, that subvariant covers the LAs of all the languages in all the datasets except those appearing in AmericasNLI, so that all AmericasNLI languages are effectively 'unseen' at fine-tuning. The second subvariant, termed LEAVE-OUT-TARG, leaves out only one language at a time from the corresponding dataset: e.g., when evaluating on Guarani (GN) in AmericasNLI, the only language 'unseen' by the TA at fine-tuning is GN as the current inference language.

The results, summarized in Tables 2(a)-(c), reveal that our MULTILINGUAL TA fine-tuning indeed increases transfer capability also for the 'TA-unseen' languages, and leads towards language-universal TAs. The scores with both subvariants offer substantial gains over MAD-X for many languages unseen during fine-tuning and in all three tasks. This confirms that (i) MAD-X TAs tend to overfit to the source language and thus underperform in cross-lingual transfer, and (ii) such overfitting might get mitigated through our proposed 'multilingual regularization' of the TAs while keeping the same modularity benefits. Additionally, the results also confirm the versatility of the proposed TLR framework, where strong transfer gains are achieved with different sets of languages included in multilingual TA fine-tuning: e.g., the scores with the two LEAVE-OUT subvariants remain strong and competitive with the full ALL-MULTI variant.

For the DP task we even observe slight gains with the LEAVE-OUT-TASK variant over the original ALL-MULTI variant which 'sees' all task languages. We speculate that this might partially occur due to the phenomenon of 'the curse of multilinguality' (Conneau et al., 2020) kicking in, now at the level of the limited TA budget, but leave this for further exploration in future work.

4.1 Further Analyses

Robustness to LA Training Configuration. To demonstrate that our results hold even when LAs are trained with the different hyper-parameters, we adopt a training regime that makes MAD-X LAs

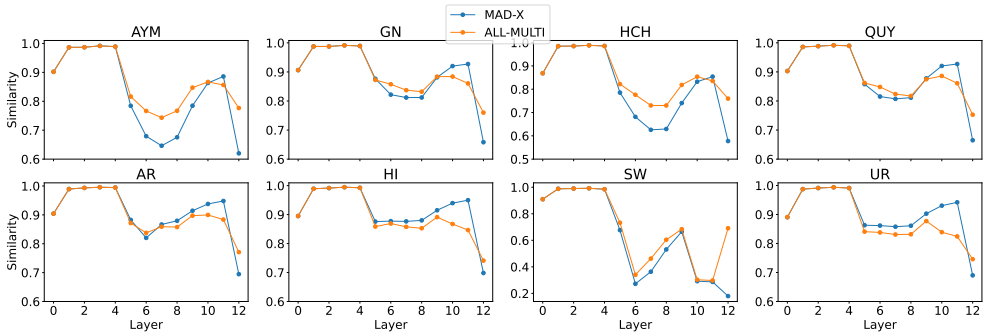

Figure 3: Per-layer similarity scores of MAD-X and ALL-MULTI adapter's representations between English and 4 languages from AmericasNLI (AYM, GN, HCH, QUY) and 4 languages from XNLI (AR, HI, SW, UR).

directly comparable with BAD-X as trained in previous work by Parović et al. (2022). The average results with such LAs on DP and AmericasNLI are presented in Table 3, demonstrating that the gains with the proposed TLR variants hold irrespective of the LA training setup.

Multiple Runs. Given the large number of experimental runs in this work, most scores are reported from single runs with fixed seeds. However, to validate that our findings hold under different random initializations of TAs, we also run MAD-X and all TLR variants with three different random seeds on a subset of tasks (MasakhaNER and AmericasNLI). The main results are presented in Table 3, indicating that all the findings hold and are not due to a single favorable seed.

Do TLR Adapters Improve Alignment Between Source and Target Languages? In order to explain the consistent gains with TLR adapters over MAD-X, we analyse whether TLR adapters produce better-aligned representations between source and target languages than MAD-X. We execute experiments on the NLI task, choosing 4 languages from AmericasNLI (AYM, GN, HCH, QUY) and 4 languages from XNLI (AR, HI, SW, UR) datasets, with English as a source language. The representations of English are obtained using MultiNLI data and English LA is paired with 1) MAD-X TA for the MAD-X baseline, and 2) ALL-MULTI TA for the TLR representations. To obtain the representations in the target language, we use its validation data and its LA paired with either MAD-X TA or ALL-MULTI TA as before. The alignment scores of both MAD-X and TLR methods are measured as cosine similarity between English and target representations of mBERT's $[CLS]$ token, using 500 examples in both languages. The results are presented in Figure 3. We can observe that MAD-X seems to have a much more significant drop in alignment values in the last layer than the ALL-MULTI adapter, which could explain the better performance of the latter. In addition, on AmericasNLI languages, where we observe sizable gains, the ALL-MULTI adapter seems to achieve better alignment across the middle layers of mBERT.

5 Related Work

Parameter-Efficient Fine-Tuning has emerged from an effort to overcome the need for full model fine-tuning, especially with the neural models becoming increasingly larger. Some approaches fine-tune only a subset of model parameters while keeping the rest unmodified (Ben Zaken et al., 2022; Guo et al., 2021; Ansell et al., 2022). Other approaches keep the model's parameters fixed and introduce a fresh set of parameters that serves for learning the desired task (Li and Liang, 2021; Lester et al., 2021; Houlsby et al., 2019; Hu et al., 2022), with the tendency towards decreasing the number of newly introduced parameters while concurrently maximizing or maintaining task performance (Karimi Mahabadi et al., 2021a,b).

Adapters were introduced in computer vision research (Rebuffi et al., 2017) before being brought into NLP to perform parameter-efficient transfer learning across tasks (Houlsby et al., 2019). Bapna and Firat (2019) use adapters in NMT as an efficient way of adapting the model to new languages and domains because maintaining separate models would quickly become infeasible as the number of domains and languages increases. Wang et al. (2021) propose factual and linguistic adapters to infuse different types of knowledge into the model, while overcoming the catastrophic forgetting that

would otherwise occur.

Adapters for Cross-Lingual Transfer. MAD-X Pfeiffer et al. (2020b) introduces LAs and TAs for efficient transfer; they also propose invertible adapters for adapting MMTs to unseen languages. Subsequently, Pfeiffer et al. (2021b) introduce a vocabulary adaptation method for MAD-X that can adapt the model to low-resource languages and even to unseen scripts, the latter of which was not possible with MAD-X's invertible adapters. In another adapter-based cross-lingual transfer approach, Vidoni et al. (2020) introduce orthogonal LAs and TAs designed to store the knowledge orthogonal to the knowledge already encoded within MMT. FAD-X (Lee et al., 2022) explores whether the available adapters can be composed to complement or completely replace the adapters for low-resource languages. This is done through fusing (Pfeiffer et al., 2021a) TAs trained with LAs in different languages. Our TLR adapters do not involve any fusion, but rather benefit from a training procedure that operates by cycling over multiple LAs. Faisal and Anastasopoulos (2022) use linguistic and phylogenetic information to improve cross-lingual transfer by leveraging closely related languages and learning language family adapters similar to Chronopoulou et al. (2022). This is accomplished by creating a phylogeny-informed tree hierarchy over LAs.

UDapter (Üstün et al., 2020) and MAD-G (Ansell et al., 2021) learn to generate LAs through the contextual parameter generation method (Platanios et al., 2018). Both UDapter and MAD-G enable the generation of the parameters from vectors of typological features through sharing of linguistic information, with the main difference between the two approaches being that MAD-G's LAs are task-agnostic, while UDapter generates them jointly with a dependency parser's parameters. Hyper-X (Üstün et al., 2022b) generates weights for adapters conditioned on both task and language vectors, thus facilitating the zero-shot transfer to unseen languages and task-language combinations.

Improving Cross-Lingual Transfer via Exposing Target Languages. In an extensive transfer case study focused on POS tagging, de Vries et al. (2022) showed that both source and target language (and other features such as language family, writing system, word order and lexical-phonetic distance) affect cross-lingual transfer performance. XeroAlign (Gritta and Iacobacci, 2021) is a method for task-specific alignment of sentence embeddings (i.e. they encourage the alignment between source task-data and its target translation by an auxiliary loss), aiming to bring the target language performance closer to that of a source language (i.e. to close the cross-lingual transfer gap). Kulshreshtha et al. (2020) analyze the effects of the existing methods for aligning multilingual contextualized embeddings and cross-lingual supervision, and propose a novel alignment method. Yang et al. (2021) introduce a new pretraining task to align static embeddings and multilingual contextual representations by relying on bilingual word pairs during masking. Inspired by this line of research, in this work we investigated how 'exposing' target languages as well as conducting multilingual fine-tuning impacts the knowledge stored in task adapters, and their ability to boost adapter-based cross-lingual transfer.

6 Conclusion and Future Work

We have presented a novel general framework for adapter-based cross-lingual task transfer, which improves over previous established adapter-based transfer frameworks such as MAD-X and BAD-X. The main idea is to better equip task adapters (TAs) to handle text instances in a variety of target languages. We have demonstrated that this can be achieved via so-called *target language-ready* (TLR) task adapters, where we expose the TA to the target language as early as the fine-tuning stage. As another major contribution, we have also proposed a multilingual language-universal TLR TA variant which offers the best trade-off between transfer performance and modularity, learning a single universal TA that can be applied over multiple target languages. Our experiments across 6 standard cross-lingual benchmarks spanning 4 different tasks and a wide spectrum of languages have validated the considerable benefits of the proposed framework and different transfer variants emerging from it. Crucially, the most modular multilingual TLR TA variant offers the strongest performance overall, and it also generalizes well even to target languages 'unseen' during TA fine-tuning.

In future work, we plan to further investigate multilingual language-universal task adapters also in multi-task and multi-domain setups, and extend the focus from serial adapters to other adapter architectures, such as parallel adapters (He et al., 2022) and sparse subnetworks (Ansell et al., 2022; Foroutan et al., 2022).

Limitations

Our experiments are based on (arguably) the most standard adapter architecture for adapter-based cross-lingual transfer and beyond, which also facilitates comparisons to prior work in this area. However, we again note that there are other emerging parameter-efficient modular methods, including different adapter architectures (He et al., 2022), that could be used with the same conceptual idea. We leave further and wider explorations along this direction for future work.

Our evaluation relies on the currently available standard multilingual benchmarks, and in particular those targeted towards low-resource languages. While the development of better models for underrepresented languages is possible mostly owing to such benchmarks, it is also inherently constrained by their quality and availability. Even though our experiments have been conducted over 35 different target languages and across several different tasks, we mostly focus on generally consistent trends across multiple languages. Delving deeper into finer-grained qualitative and linguistically oriented analyses over particular low-resource languages would require access to native speakers of those languages, and it is very challenging to conduct such analyses for many languages in our language sample.

Due to a large number of experiments across many tasks and languages, we report all our results based on a single run. Averages over multiple runs conducted on a subset of languages and tasks confirm all the core findings; for simplicity, we eventually chose to report the results for all languages and tasks in the same setup.

Finally, training language adapters is typically computationally expensive; however, owing to the modular design of our framework with respect to language adapters, these are trained only once per language and reused across different evaluations.

Acknowledgments

We would like to thank the reviewers for their helpful suggestions.

Marinela Parović is supported by Trinity College External Research Studentship. Alan wishes to thank David and Claudia Harding for their generous support via the Harding Distinguished Postgraduate Scholarship Programme. Ivan Vulić is supported by a personal Royal Society University Research Fellowship *'Inclusive and Sustainable Language Technology for a Truly Multilingual World'* (no 221137; 2022–).

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

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.

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

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. 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 1538–

1548, Hong Kong, China. Association for Computational Linguistics.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Dublin, Ireland. Association for Computational Linguistics.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2022. Language-family adapters for multilingual neural machine translation. *CoRR*, abs/2209.15236.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

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.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.

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

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

Fahim Faisal and Antonios Anastasopoulos. 2022. Phylogeny-inspired adaptation of multilingual models to new languages. In *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 434–452, Online only. Association for Computational Linguistics.

Negar Foroutan, Mohammadreza Banaei, Rémi Lebret, Antoine Bosselut, and Karl Aberer. 2022. Discovering language-neutral sub-networks in multilingual language models. *CoRR*, abs/2205.12672.

Milan Gritta and Ignacio Iacobacci. 2021. XeroAlign: Zero-shot cross-lingual transformer alignment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 371–381, Online. Association for Computational Linguistics.

Demi Guo, Alexander Rush, and Yoon Kim. 2021. Parameter-efficient transfer learning with diff pruning. 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 4884–4896, Online. Association for Computational Linguistics.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a unified view of parameter-efficient transfer learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International*

Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

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

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: an empirical study. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. 2021a. Compacter: Efficient low-rank hypercomplex adapter layers. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 1022–1035.

Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. 2021b. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. 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 565–576, Online. Association for Computational Linguistics.

Saurabh Kulshreshtha, Jose Luis Redondo Garcia, and Ching-Yun Chang. 2020. Cross-lingual alignment methods for multilingual BERT: A comparative study. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 933–942, Online. Association for Computational Linguistics.

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

Jaeseong Lee, Seung-won Hwang, and Taesup Kim. 2022. FAD-X: Fusing adapters for cross-lingual transfer to low-resource languages. In *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 2: Short Papers)*, pages 57–64, Online only. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. 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, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021a. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. 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.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021b. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jerin Philip, Alexandre Berard, Matthias Gallé, and Laurent Besacier. 2020. Monolingual adapters for zero-shot neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4465–4470, Online. Association for Computational Linguistics.

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

Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018. Contextual parameter generation for universal neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 425–435, Brussels, Belgium. Association for Computational Linguistics.

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

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

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

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.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2022a. UDapter: Typology-based language adapters for multilingual dependency parsing and sequence labeling. *Computational Linguistics*, 48(3):555–592.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, Gertjan van Noord, and Sebastian Ruder. 2022b. Hyper-x: A unified hypernetwork for multi-task multilingual transfer. *arXiv preprint arXiv:2205.12148*.

Marko Vidoni, Ivan Vulić, and Goran Glavaš. 2020. Orthogonal language and task adapters in zero-shot cross-lingual transfer. *arXiv preprint arXiv:2012.06460*.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418, Online. Association for Computational Linguistics.

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

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Ziqing Yang, Wentao Ma, Yiming Cui, Jiani Ye, Wanxiang Che, and Shijin Wang. 2021. Bilingual alignment pre-training for zero-shot cross-lingual transfer. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 100–105, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, et al. 2020. Universal dependencies 2.7. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A Tasks and Languages

The full list of tasks, datasets and target languages with their names and codes is given in Table 5.

B XLM-R Results

The results on AmericasNLI, XNLI and XQuAD with XLM-R are shown in Table 6.

C MAD-X Adapters Trained with a Different Setup

The results of MAD-X adapters trained in a different setup (Parović et al., 2022) on DP and AmericasNLI are given in Table 7. The results of these adapters are directly comparable with the BAD-X baseline, as they follow the same training setup and their summary is given in Table 3.

D Per-Language Results with Multiple Runs

Full results on MasakhaNER and AmericasNLI for all target languages obtained as an average across 3 different random seeds are given in Table 8.

Task	Source Dataset	Target Dataset	Target Languages
Dependency Parsing (DP)	Universal Dependencies 2.7 (Zeman et al., 2020)	Universal Dependencies 2.7 (Zeman et al., 2020)	Afrikaans (AF)*, Bambara (BM), Basque (EU)*, Komi-Zyryan (KPV), Marathi (MR)*, Maltese (MT), Erzya (MYV), Telugu (TE)*, Uyghur (UG), Wolof (WO)
Named Entity Recognition (NER)	CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003)	MasakhaNER (Adelani et al., 2021)	Hausa (HAU), Igbo (IBO), Kinyarwanda (KIN), Luganda (LUG), Luo (LUO), Nigerian-Pidgin (PCM), Swahili (SWA)*, Wolof (WOL), Yorùbá (YOR)*
Natural Language Inference (NLI)	MultiNLI (Williams et al., 2018)	AmericasNLI (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 (Williams et al., 2018)	XNLI (Conneau et al., 2018)	Arabic (AR)†, Hindi (HI)†, Swahili (SW)*, Thai (TH)†, Urdu (UR)*, Chinese (ZH)†
Question Answering (QA)	SQuAD v1.1 (Rajpurkar et al., 2016)	XQuAD (Artetxe et al., 2020)	Arabic (AR)†, Hindi (HI)†, Thai (TH)†, Chinese (ZH)†
	SQuAD v1.1 (Rajpurkar et al., 2016)	TyDiQA-GoldP (Clark et al., 2020)	Arabic (AR)†, Bengali (BN)*, Swahili (SW)*, Telugu (TE)*, Thai (TH)†

Table 5: Details of the tasks, datasets, and languages involved in our cross-lingual transfer evaluation. * denotes low-resource languages seen during MMT pretraining; † denotes high-resource languages seen during MMT pretraining; all other languages are low-resource and unseen. The source language is always English.

Method	AYM	BZD	CNI	GN	HCH	NAH	OTO	QUY	SHP	TAR	avg	Better
MAD-X	54.40	40.40	46.80	58.13	40.80	48.92	44.39	55.47	50.67	42.53	48.25	
TARGET	52.67	43.73	46.13	58.93	44.80	49.59	43.45	57.47	48.67	41.87	48.73	5/10
BILINGUAL	53.47	43.47	47.20	58.40	44.40	49.73	41.98	57.73	47.87	42.27	48.65	6/10
TASK-MULTI	53.20	43.73	47.47	56.67	42.27	49.59	42.51	58.67	48.93	43.73	48.68	6/10
ALL-MULTI	53.47	42.27	47.73	57.47	41.47	49.73	40.91	58.80	50.27	40.93	48.31	5/10

(a) AmericasNLI: accuracy

Method	AR	HI	SW	TH	UR	ZH	avg	Better
MAD-X	66.81	63.89	64.83	63.41	60.76	67.43	64.52	
TARGET	67.19	66.37	63.99	67.05	61.84	70.40	66.14	5/6
BILINGUAL	66.67	66.07	64.37	66.67	61.68	70.04	65.92	4/6
TASK-MULTI	68.00	65.89	64.19	66.01	61.30	69.58	65.83	5/6
ALL-MULTI	67.84	66.11	64.89	65.67	61.82	69.34	65.95	6/6

(b) XNLI: accuracy

Method	AR	HI	TH	ZH	avg	Better
MAD-X	65.23/47.65	67.15/51.09	69.26/59.08	64.01/55.13	66.41/53.24	
TARGET	65.63/48.40	69.49/53.78	69.38/58.57	64.09/54.71	67.15/53.87	4/4
BILINGUAL	65.85/48.91	68.27/52.86	70.31/60.50	64.57/55.55	67.25/54.45	4/4
TASK-MULTI	66.23/48.40	68.43/52.61	70.25/60.42	65.32/56.22	67.56/54.41	4/4
ALL-MULTI	65.98/49.24	68.24/51.60	67.15/56.55	63.07/52.94	66.11/52.58	2/4

(c) XQuAD: F1/EM

Table 6: XLM-R: Results of all methods and TLR variants on all target languages.

Method	AF	BM	EU	KPV	MR	MT	MYV	TE	UG	WO	avg	Better
MAD-X	54.23	11.80	32.51	22.44	24.24	44.71	35.45	45.47	15.67	26.38	31.29	
BAD-X	54.54	11.92	31.45	22.55	26.56	43.52	39.31	46.22	15.24	35.28	32.66	
TARGET	55.07	11.96	33.31	20.82	28.05	48.83	41.75	**52.34**	18.60	40.75	35.15	9/10
BILINGUAL	54.75	11.86	33.21	22.09	26.60	48.74	38.82	49.86	16.89	41.27	34.41	9/10
TASK-MULTI	**56.55**	11.94	34.17	23.82	27.71	51.66	40.87	51.10	**18.90**	41.93	35.86	10/10
ALL-MULTI	56.28	**12.91**	**35.04**	**24.11**	**28.28**	**53.02**	**41.85**	51.43	18.47	**43.31**	**36.47**	10/10

(a) DP: LAS

Method	AYM	BZD	CNI	GN	HCH	NAH	OTO	QUY	SHP	TAR	avg	Better
MAD-X	47.07	**45.07**	41.87	55.33	39.47	48.51	40.91	51.47	41.60	42.00	45.33	
BAD-X	46.13	44.67	45.87	56.80	44.93	47.70	41.71	47.87	**49.07**	39.47	46.42	
TARGET	48.80	44.80	44.13	58.27	43.73	**51.90**	41.84	57.47	46.40	45.07	48.24	9/10
BILINGUAL	**49.87**	44.13	45.87	60.40	43.47	50.27	41.98	**58.00**	46.53	44.13	48.47	9/10
TASK-MULTI	46.40	44.27	45.87	57.60	44.40	50.68	42.78	**58.00**	46.53	44.00	48.05	8/10
ALL-MULTI	46.00	44.00	**46.40**	**61.07**	**46.53**	49.32	**44.12**	55.33	46.67	**45.47**	**48.49**	8/10

(b) AmericasNLI: accuracy

Table 7: Results of all methods and TLR variants on DP and AmericasNLI across all target languages. All adapters in these experiments have been trained using the hyperparameters from Parović et al. (2022). The highest task score per each language is in **bold**. *Better* refers to the number of target languages for which each TLR variant scores higher than MAD-X.

Method	HAU	IBO	KIN	LUG	LUO	PCM	SWA	WOL	YOR	avg	Better
MAD-X	**82.00**	70.92	63.55	65.26	48.62	72.40	74.53	64.35	72.78	68.27	
TARGET	78.32	71.70	63.35	67.52	**50.88**	73.99	75.46	62.55	72.68	68.49	5/9
BILINGUAL	80.68	71.56	63.92	**68.11**	50.49	**74.78**	**76.43**	64.39	**72.80**	69.24	8/9
TASK-MULTI	81.85	**72.18**	**65.39**	66.98	50.61	74.42	76.14	**65.58**	72.07	**69.47**	7/9
ALL-MULTI	81.49	71.32	64.86	66.26	50.68	74.42	75.70	65.52	71.66	69.10	7/9
LEAVE-OUT-TASK	82.30	70.79	65.61	67.50	50.81	74.24	75.69	65.32	72.08	69.37	7/9
LEAVE-OUT-TARG	82.41	70.66	65.35	67.38	50.95	73.90	75.52	64.86	71.18	69.13	7/9

(a) NER: F1

Method	AYM	BZD	CNI	GN	HCH	NAH	OTO	QUY	SHP	TAR	avg	Better
MAD-X	51.55	41.24	39.47	56.62	40.09	45.98	40.82	49.29	40.71	40.84	44.66	
TARGET	50.89	**46.62**	43.42	57.20	43.42	**49.37**	41.31	56.31	46.62	44.00	47.92	9/10
BILINGUAL	**53.69**	46.18	43.60	58.40	44.31	47.92	42.96	56.00	46.98	43.20	48.32	10/10
TASK-MULTI	51.11	45.38	44.80	58.49	45.51	49.05	42.96	56.31	47.65	**44.22**	48.55	9/10
ALL-MULTI	52.62	45.69	**45.91**	**59.07**	**45.78**	48.51	**45.01**	**56.84**	**47.82**	43.78	**49.10**	10/10
LEAVE-OUT-TASK	53.91	43.60	45.78	57.87	42.80	47.56	42.87	56.40	46.13	42.66	47.96	10/10
LEAVE-OUT-TARG	52.09	44.98	45.91	58.13	44.44	48.74	44.43	56.13	46.98	42.58	48.44	10/10

(b) AmericasNLI: accuracy

Table 8: Averages across 3 different random seeds of all methods and TLR variants on MasakhaNER and Americas-NLI across all target languages. The highest task score per each language is in **bold**, but excluding the two ablation subvariants of ALL-MULTI placed below the dashed horizontal lines (LEAVE-OUT-TASK and LEAVE-OUT-TARG). *Better* refers to the number of target languages for which each TLR variant scores higher than MAD-X.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7 (Limitations)

☑ A2. Did you discuss any potential risks of your work?
7 (Limitations)

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract, 1 (Introduction)

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3, Appendix A

☑ B1. Did you cite the creators of artifacts you used?
3, Appendix A

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3, Appendix A

C ☑ Did you run computational experiments?

3, 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
3, 4, 7

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.