To Translate or Not to Translate: A Systematic Investigation of Translation-Based Cross-Lingual Transfer to Low-Resource Languages

Benedikt Ebing and Goran Glavaš
University of Würzburg
Center for Artificial Intelligence and Data Science (CAIDAS)
{benedikt.ebing, goran.glavas}@uni-wuerzburg.de

Abstract

Perfect machine translation (MT) would render cross-lingual transfer (XLT) by means of multilingual language models (mLMs) superfluous. Given, on the one hand, the large body of work on improving XLT with mLMs and, on the other hand, recent advances in massively multilingual MT, in this work, we systematically evaluate existing and propose new translation-based XLT approaches for transfer to low-resource languages. We show that all translation-based approaches dramatically outperform zero-shot XLT with mLMs—with the combination of round-trip translation of the source-language training data and the translation of the target-language test instances at inference—being generally the most effective. We next show that one can obtain further empirical gains by adding reliable translations to other high-resource languages to the training data. Moreover, we propose an effective translation-based XLT strategy even for languages not supported by the MT system. Finally, we show that model selection for XLT based on target-language validation data obtained with MT outperforms model selection based on the source-language data. We believe our findings warrant a broader inclusion of more robust translation-based baselines in XLT research.

1 Introduction

Multilingual language models (mLMs) like mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), or mT5 (Xue et al., 2021) have become the backbone of multilingual NLP. Their multilingual pretraining and the consequent ability to encode texts from a wide range of languages make them suitable for cross-lingual transfer (XLT) for downstream NLP tasks: fine-tuned on available task-specific data in high-resource languages, they can be used to make predictions for languages that lack task-specific (training) data. Their effectiveness as vehicles of both zero-shot (no task-specific training instances in the target language, ZS-XLT) and few-shot XLT (few training instances in the target language, FS-XLT) has been documented for a plethora of tasks and languages (Wu and Dredze, 2019; Wang et al., 2019; Lauscher et al., 2020; Schmidt et al., 2022). Cross-lingual transfer with mLMs, however, yields poor performance for low-resource target languages that are (i) un(der)represented in the pretraining corpora, especially if they are additionally (ii) linguistically distant from the source language (Lauscher et al., 2020; Adelani et al., 2022; Ebrahimi et al., 2022).

Recent years have witnessed a large body of work that focused on improving XLT, in particular for low-resource target languages. First, a multitude of new multilingual benchmarks have been introduced, aiming to either evaluate XLT with mLMs on sets of linguistically diverse languages (Clark et al., 2020; Ponti et al., 2020; Ruder et al., 2021) or on groups of related low-resource languages from underrepresented language families (i.e., families without any high-resource language) and/or geographies (Adelani et al., 2021, 2022; Ebrahimi et al., 2022; Aggarwal et al., 2022; Muhammad et al., 2022; Armstrong et al., 2022; Winata et al., 2023, *inter alia*). Second, a diverse set of methodological proposals have been introduced, ranging from (i) attempts to better align mLMs' representation subspaces of languages (Wu and Dredze, 2020; Hu et al., 2021; Yang et al., 2022, inter alia) over (ii) those that increase the representational capacity for underrepresented languages, typically via additional post-hoc language-specific language modeling training (Pfeiffer et al., 2020, 2022; Ansell et al., 2022; Parović et al., 2022, *inter alia*) to (iii) various FS-XLT proposals that seek to maximally exploit small sets of task-specific target language instances (Hedderich et al., 2020;

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5325–5344
June 16-21, 2024 ©2024 Association for Computational Linguistics

Lauscher et al., 2020; Zhao et al., 2021; Schmidt et al., 2022, *inter alia*).

Much of the above work rendered *translation-based XLT strategies*—in which an MT model is employed to either translate the source-language training data into the target language before training (referred to as *translate-train*) or translate the target-language instances to the source language before inference (*translate-test*)—competitive w.r.t. mLM-based transfer (Hu et al., 2020; Ruder et al., 2021; Ebrahimi et al., 2022; Aggarwal et al., 2022). *Sporadically*, however, MT has been leveraged for more elaborate translation-based strategies—e.g., translating source-language training data to multiple (related) target languages (Hu et al., 2020), combining the translated target-language training data with the original source-language training data (Chen et al., 2023), or using monolingual English LMs instead of mLMs for translate-test (Artetxe et al., 2020, 2023)—complicating the selection of translation-based baselines in XLT research. In fact, the most recent evidence (Artetxe et al., 2023) suggests that the potential of translation-based XLT has been underestimated due to the selection of sub-optimal translation strategies. What is more, much of the work on low-resource XLT completely disregards translation-based baselines, arguing *a priori*, without empirical confirmation, that (1) due to the lack of parallel data, MT models for low-resource languages exhibit poor performance, which directly caps the potential of translation-based XLT and/or (2) their evaluation encompasses target languages that are unsupported by (state-of-the-art, commercial) MT systems.

Two recent developments, however, warrant a systematic (re-)evaluation of translation-based XLT for low-resource languages: (i) the availability of open massively multilingual MT models that not only support an increasingly large set of languages (Tiedemann and Thottingal, 2020; Liu et al., 2020; Fan et al., 2021; Team et al., 2022; Kudugunta et al., 2023), but also yield meaningful translations even for the smallest of those languages; and (ii) recent proposals of novel translation-based XLT strategies that have been largely uninvestigated in XLT to truly low-resource languages (Hu et al., 2020; Chen et al., 2023; Artetxe et al., 2023).

Contributions. In this work, we contribute to the body of translation-based XLT in light of these recent advances, focusing explicitly on low-resource target languages: **1)** we offer a systematic comparison of different translation-based XLT strategies on three established benchmarks for sequence- and token-level classification, encompassing in total 40 different low-resource languages; **2)** Motivated by the success of multi-source training (Ruder, 2017; Ansell et al., 2021) and ensembling (Oh et al., 2022), as well as the high quality of MT between high-resource languages, we propose two novel strategies that integrate translations from the source data to three diverse high-resource languages (Turkish, Russian, and Chinese); we find that integrating translations to other high-resource languages substantially improves performance for sequence-level classification tasks; **3)** We propose a simple and effective translation-based XLT approach for languages *not covered* by the MT models in which we translate from/to the linguistically closest supported language, demonstrating substantial gains over ZS-XLT with mLMs; **4)** We introduce a translation-based *model selection* in which the optimal model checkpoint is selected based on performance on the validation data automatically translated to the target language; we show that this results in better performance than model selection based on source-language validation data. **5)** Finally, we run several ablations, offering insights into the impact of lower-level design decisions—such as the MT decoding strategy or joint vs. sequential fine-tuning—on translation-based XLT.

2 Translation-Based Strategies

Most of the existing XLT work evaluates only the most straightforward *translate-train* (`T-Train`) and *translate-test* (`T-Test`) baselines. The former assumes the translation of the training data, available in some high-resource language (almost always English), to the target language in which inference is performed. The latter trains on the clean source-language data but, at inference time, translates the target language instances to the source language before making predictions. More recent works (Oh et al., 2022; Artetxe et al., 2023) propose a combination of the two, which we dub *roundtrip-train-test* (`RTT`), where the source-language training data is round-trip translated (i.e., to the target language and then back) so that the translated test data at inference time better matches the training distribution, reflecting the idiosyncrasies of the same MT model. In what follows, we describe the variants of

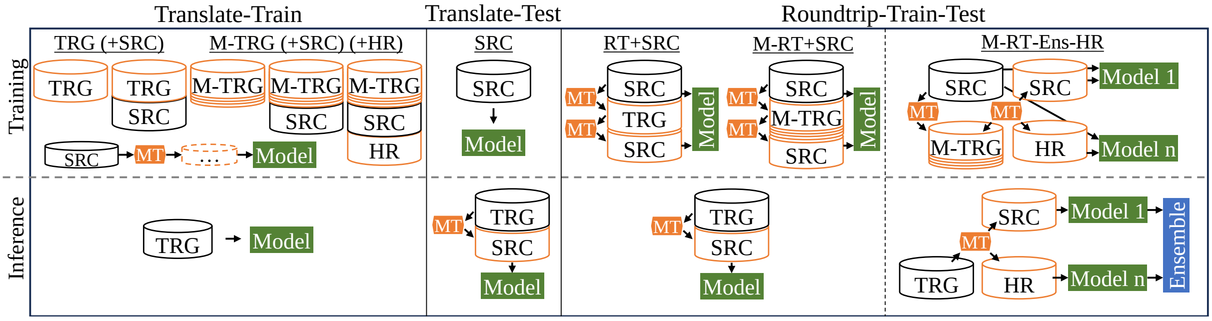

Figure 1: Schematic overview of translation-based XLT methods. Clean source or target language data is indicated in black, while noisy translated data is shown in orange.

T-Train, T-Test, and RTT that we evaluate. Figure 1 concisely illustrates all MT-based approaches under evaluation.

2.1 Translate-Train (T-Train)

Target (TRG). This is the standard T-Train where the source-language training data is translated into one particular target language. The mLM is then fine-tuned on the automatically translated (i.e., noisy) target-language training dataset.

Multi-Target (M-TRG). This is a generalization of T-Train in which we translate the source-language training data into each language from a set of (presumably related) target languages: in our experiments, these are all languages of a particular benchmark dataset supported by the MT model, e.g., all AmericasNLI languages (Ebrahimi et al., 2022). We then fine-tune the mLM in a multi-source setup, i.e., on the concatenation of the training data translated to each of the target languages (per task).

Keeping the Source-Language Data (+SRC). In this variant, we concatenate the noisy translated training dataset in the target language (or a set of target languages) with the original (i.e., clean) training data in the source language. We denote these variants TRG+SRC (if we concatenate source language data to TRG) and M-TRG+SRC (if we concatenate the source-language data to M-TRG).

Adding Diverse High-Resource Languages (+HR). We additionally explore translating the source-language training data to a (small) set of linguistically diverse high-resource languages. The motivation for this is two-fold: (1) multilingual (i.e., multi-source) fine-tuning has been shown to bring benefits compared to monolingual (English-only) fine-tuning (Ansell et al., 2021); and (2) automatic translation from high-resource source language (i.e., English) to other high-resource languages (i.e., Chinese, Turkish, and Russian) is generally of much higher quality than translation to low-resource target languages (e.g., Guarani). Exploiting strong MT between high-resource languages will, under this assumption, allow us to obtain linguistically diverse yet high-quality training data, which should consequently lead to improvements in XLT to any low-resource language. We evaluate variants in which translations to high-resource languages are added to TRG+SRC (i.e., TRG+SRC+HR) and M-TRG+SRC (i.e., M-TRG+SRC+HR).

2.2 Translate-Test (T-Test)

We evaluate the standard T-Test baseline where the model is trained on the original source-language data and, at inference time, the target-language instances are translated to the source language before the source-language model makes the prediction.

2.3 Roundtrip-Train-Test (RTT)

Round-Trip T-Train + T-Test (RT). Prior work suggested that the mismatch between high-quality training data and noisily translated evaluation data poses a challenge for the T-Test approach (Artetxe et al., 2020; Oh et al., 2022; Artetxe et al., 2023). To overcome this shift in data distribution that the model is exposed to at test time, in RTT, we also train on the noisy source-language data obtained via round-trip translation of the original clean source-language data to the target language and back. Similar to T-Train, we evaluate the variants of RTT where the noisy source-language data is obtained via round-trip translation to a single target language (denoted with RT) and multiple target languages (M-RT) and, finally, concatenated to the original (i.e., clean) source-language data (RT+SRC, M-RT+SRC).

Model Ensembling for RTT (M-RT-Ens). Following our idea of exploiting other high-resource languages in translation-based XLT, we propose a novel RTT variant in which we not only round-trip translate the source-language data into the target language and back into the source language but also translate the source data into the target language and then into different high-resource languages, other than the initial source language (e.g., Source→Target→Chinese). We apply this paradigm to the same three high-resource languages used for the T-Train-based approaches (i.e., Chinese, Russian, and Turkish). Here in ensembling, however, for each of these high-resource languages, we independently fine-tune an mLM instance on the round-trip translated data of that language, concatenated with the original source-language (i.e., English) data (e.g., for English as source and Chinese as the high-resource auxiliary, we concatenate the clean original English with the noisy Chinese data obtained via two-step translation). Finally, we ensemble the predictions of the (four) fine-tuned models (English, Chinese, Turkish, Russian): we average the class probability distributions of the models obtained for a target-language test instance, previously translated to each of the high-resource languages, respectively. We denote this RTT ensemble approach with M-RT-Ens-HR. Since model ensembles are known to outperform single models (Wortsman et al., 2022), in our experiments, we compare M-RT-Ens-HR against the ensemble (of equally many models) fine-tuned on the round-trip translated source-language data (i.e., round-trip translated English) only, using different random seeds (we denote this with M-RT-Ens-SRC).

2.4 Unsupported Languages

Even though recent multilingual MT models cover a broad range of low-resource languages, the majority of the world's languages remain unsupported. Motivated by prior work on finding the best transfer source for a given target language (Lin et al., 2019; Adelani et al., 2022; Glavaš and Vulić, 2021), we propose to translate to (T-Train) and from (T-Test) an MT-supported language that is linguistically closest to the unsupported target: to this end, we quantify the linguistic proximity of languages as the cosine similarity of their typological vectors from the URIEL database (Littell et al., 2017).

3 Experimental Setup

Machine Translation. For translation, we leverage the state-of-the-art massively multilingual NLLB model with 3.3B parameters (Team et al., 2022). Building on prior work (Artetxe et al., 2023), we ablate over decoding strategies, including greedy decoding, nucleus sampling with top-p = 0.8, and beam search with beam size 5. In our final evaluation, translations are generated using beam search.

Evaluation Tasks and Datasets.. Following prior work on low-resource XLT (Ansell et al., 2021, 2022; Schmidt et al., 2022), we evaluate on sequence- and token-level classification tasks covering languages un(der)represented in the pretraining corpus of our base models. In all experiments, English is the source language.[1]

Natural Language Inference (NLI). We evaluate our approaches on AmericasNLI (AmNLI) (Ebrahimi et al., 2022). AmNLI contains 10 indigenous languages of the Americas, only 3 of which are supported by the NLLB model we use and none are present in the pretraining corpus of our backbone model. We utilize the English training and validation portion of XNLI (Conneau et al., 2018) as our source-language data. The dataset covers 393k training and 2490 validation instances. We jointly encode the hypothesis-premise pair and feed the transformed sequence start token into a feed-forward softmax classifier.

Text Classification (TC). We use the sentiment classification dataset NusaX (Winata et al., 2023), which comprises 10 languages from Indonesia, 7 of which are supported by the NLLB model and 2 are seen by our backbone model in pretraining. The English training (500 instances) and validation portions (100 instances) are used as our source-language data. Similar to NLI, we feed the transformed representation of the sequence start token—output of the Transformer encoder—into the softmax classifier.

Named Entity Recognition (NER). Our evaluation spans a set of 20 languages from MasakhaNER 2.0

[1] We provide the complete list of languages in App. A.

(Masakha) (Adelani et al., 2022). The dataset comprises a diverse set of underrepresented languages spoken in Sub-Saharan Africa. Among these, 18 languages are supported by the NLLB model we use for MT, but only 3 are covered in the pretraining corpus of our backbone model. Our source data are the English training and validation portions of CoNLL (Tjong Kim Sang and De Meulder, 2003), with more than 14k instances for training and 3250 validation instances. In this token-level task, the classifier makes a prediction from the output (i.e., transformed) representation of each input token.

Word Aligner. Translation-based transfer for token-level tasks requires *label projection*, i.e., mapping of the labels from source-language tokens to the tokens of the translated target sequence. To that end, we map labels post-translation with AccAlign (Wang et al., 2022), a state-of-the-art word aligner based on the multilingual sentence encoder LaBSE (Feng et al., 2022).[2] When recovering the labels for the translated sequences, we discard a training instance whenever we cannot map a labeled source-language token to its target-language counterpart. The projection rates (for training data), i.e., the percentage of successful token mappings, is given for all supported languages in the App. B.

Downstream Fine-Tuning. We use XLM-R (Large) (Conneau et al., 2020) in all our experiments. For T-Test and RTT, we also experiment with RoBERTa (Large) (Liu et al., 2019). We outline the downstream fine-tuning details in Appendix C. We evaluate models at various checkpoints of training: (i) at the end of the epoch[3] with the best performance on source-language validation data (*Val-Src*), (ii) at the end of the epoch with the best performance on source-language validation data machine translated to the target language (*Val-MT-Trg*), and (iii) at the end of the epoch with the best performance on target-language validation data (*Val-Trg*). *Val-MT-Trg* and *Val-Trg* cannot be directly applied to T-Test and RTT as both model selection methods use (translated) target language data, while the training data of T-Test and RTT is solely in English. Hence, we adapt *Val-MT-Trg* and *Val-Trg* for T-Test and RTT: for *Val-MT-Trg*, we conduct round-trip translation on the source validation data pivoting through the target language (i.e., Source→Target→Source), and for *Val-Trg*, we simply MT-ed the (oracle) target validation data to the source language. Unless specified otherwise, we report results based on *Val-Src* and show the results for *Val-MT-Trg* and *Val-Trg* in Appendix E. We run experiments with 3 distinct random seeds and report mean accuracy for NLI and average F1 for NER and TC, as well as the standard deviation.

		AmNLI	NusaX	Masakha	Avg
Zero-Shot					
SRC	X	$44.7_{\pm1.2}$	$71.2_{\pm1.3}$	$47.9_{\pm0.6}$	$54.6_{\pm1.1}$
Translate-Train					
TRG	X	$61.1_{\pm0.4}$	$77.8_{\pm0.8}$	$62.1_{\pm0.3}$	$67.0_{\pm0.5}$
TRG+SRC	X	$62.4_{\pm0.3}$	$79.7_{\pm0.6}$	$\mathbf{64.1}_{\pm0.3}$	$\mathbf{68.8}_{\pm0.4}$
M-TRG	X	$63.4_{\pm0.5}$	$79.0_{\pm0.7}$	$56.9_{\pm0.4}$	$66.4_{\pm0.5}$
M-TRG+SRC	X	$63.6_{\pm0.6}$	$80.8_{\pm0.4}$	$57.4_{\pm0.6}$	$67.3_{\pm0.5}$
incl. Translations to High-Resource Languages					
TRG+SRC+HR	X	$62.9_{\pm0.5}$	$78.1_{\pm1.3}$	$62.9_{\pm0.3}$	$68.0_{\pm0.8}$
M-TRG+SRC+HR	X	$64.7_{\pm0.4}$	$79.1_{\pm1.9}$	$58.0_{\pm0.5}$	$67.3_{\pm1.2}$
Translate-Test					
SRC	R	$53.1_{\pm0.1}$	$79.4_{\pm0.4}$	$54.7_{\pm0.1}$	$62.4_{\pm0.2}$
SRC	X	$52.9_{\pm0.5}$	$80.9_{\pm0.8}$	$54.1_{\pm0.1}$	$62.6_{\pm0.5}$
Roundtrip-Train-Test					
RT+SRC	R	$62.4_{\pm0.6}$	$81.2_{\pm0.4}$	$54.6_{\pm0.1}$	$66.1_{\pm0.4}$
RT+SRC	X	$63.1_{\pm0.4}$	$81.6_{\pm0.5}$	$53.6_{\pm0.2}$	$66.1_{\pm0.4}$
M-RT+SRC	R	$64.3_{\pm0.2}$	$81.0_{\pm0.4}$	$54.0_{\pm0.2}$	$66.4_{\pm0.3}$
M-RT+SRC	X	$64.0_{\pm0.3}$	$82.1_{\pm0.4}$	$53.0_{\pm0.4}$	$66.4_{\pm0.4}$
M-RT-Ens-SRC	X	$63.7_{\pm0.2}$	$82.8_{\pm0.3}$	$53.7_{\pm0.1}$	$66.7_{\pm0.2}$
incl. Translations to High-Resource Languages					
M-RT-Ens-HR	X	$\mathbf{66.1}_{\pm0.2}$	$\mathbf{83.9}_{\pm0.4}$	$45.8_{\pm0.1}$	$65.3_{\pm0.3}$

Table 1: Results for translation-based XLT for languages supported by the MT model. We use XLM-R (X) and RoBERTa (R). The best results are shown in **bold**.

4 Main Results and Discussion

Table 1 summarizes our main results: performance of MT-based T-Train, T-Test, and RTT variants in low-resource XLT on three low-resource XLT benchmarks.

T-Train vs. T-Test. We first assess the widely used T-Train and T-Test baselines. These simple translation-based XLT strategies outperform ZS-XLT dramatically: from 6.2% on Masakha (T-Test with XLM-R) up to 18.9% on AmNLI (M-TRG+SRC), rendering them as unavoidable baselines for any XLT effort. Keeping the original clean source language data in the training mix is beneficial: TRG+SRC and M-TRG+SRC consistently outperform TRG and M-TRG, respectively. For sequence-level classification tasks (AmNLI and NusaX),

[2] We adhere to the hyperparameters specified in their work.
[3] For AmNLI, we checkpoint after every 10% of an epoch.

training on the concatenation of the clean source data and the source data translated to a set of related target languages (M-TRG+SRC) yields the best results. For NER on Masakha, TRG+SRC maximizes XLT performance. We further observe that the optimal T-Train (TRG+SRC) strategy significantly outperforms (+6.2%) the best T-Test approach. Our T-Test results also demonstrate that in low-resource XLT, mLMs yield comparable performance to monolingual LMs: this contradicts the recent T-Test finding for high-resource languages of Artetxe et al. (2023).

RTT. For sequence-level classification tasks, we find that RTT outperforms the best T-Train strategy (M-TRG+SRC), which is in line with prior findings (Artetxe et al., 2023; Oh et al., 2022). For NusaX, this observation holds for all RTT variants. For AmNLI, only M-RT+SRC consistently outperforms M-TRG+SRC. We further observe inconclusive results regarding the LM for which we get the highest performance for M-RT+SRC: while RoBERTa is superior on AmNLI, XLM-R displays better performance on NusaX. This result, however, does not extend to RT+SRC, for which XLM-R consistently outperforms RoBERTa. As already seen, T-Test lags T-Train on Masakha, and this is also true for RTT. Even more so, RTT progressively degrades in performance the more round-trip translated data is introduced (i.e., RT+SRC trails T-Test by at least 0.1% whereas M-RT+SRC does so by 0.7%). We hypothesize that both the amount of round-trip translated data and the type of task drive the performance of monolingual LMs like RoBERTa in translation-based XLT to low-resource languages. Our results challenge prior work (Artetxe et al., 2023; Oh et al., 2022), in which T-Test and RTT are better with monolingual LMs than with mLMs. Their experiments, however, covered predominantly high-resource target languages.

Adding High-Resource Languages. Table 1 further reports results of T-Train and RTT variants that include high-resource languages (i.e., Chinese, Russian, and Turkish) for translation-based XLT. The results for T-Train are inconsistent. For AmNLI, including high-resource languages (M-TRG+SRC+HR) boosts performance by at least 1.1%. These gains persist for different model selection strategies (cf. Appendix E). However, such multilingual data augmentation adversely affects the performance on NusaX and Masakha.

	AmNLI	NusaX	Masakha	Avg
Translate-Train				
Val-Src	$62.6_{\pm 0.5}$	$79.3_{\pm 0.6}$	$60.1_{\pm 0.4}$	$67.3_{\pm 0.5}$
Val-MT-Trg	$62.8_{\pm 0.5}$	$79.6_{\pm 0.7}$	$60.3_{\pm 0.3}$	$67.6_{\pm 0.5}$
Val-Trg	$\underline{62.9}_{\pm 0.5}$	$\underline{80.2}_{\pm 0.6}$	$\underline{62.2}_{\pm 0.4}$	$\mathbf{68.4}_{\pm 0.5}$
Translate-Test				
Val-Src	$53.0_{\pm 0.4}$	$80.1_{\pm 0.6}$	$\underline{54.4}_{\pm 0.6}$	$62.5_{\pm 0.5}$
Val-MT-Trg	$53.1_{\pm 0.5}$	$79.8_{\pm 0.5}$	$54.3_{\pm 0.1}$	$62.4_{\pm 0.4}$
Val-Trg	$\underline{53.4}_{\pm 0.4}$	$\underline{80.8}_{\pm 0.7}$	$\underline{54.4}_{\pm 0.1}$	$\mathbf{62.9}_{\pm 0.5}$
Roundtrip-Train-Test				
Val-Src	$\underline{63.5}_{\pm 0.4}$	$81.5_{\pm 0.4}$	$53.8_{\pm 0.3}$	$66.3_{\pm 0.4}$
Val-MT-Trg	$\underline{63.5}_{\pm 0.4}$	$81.4_{\pm 0.5}$	$53.7_{\pm 0.2}$	$66.2_{\pm 0.4}$
Val-Trg	$63.4_{\pm 0.5}$	$\underline{81.7}_{\pm 0.4}$	$\underline{54.0}_{\pm 0.2}$	$\mathbf{66.4}_{\pm 0.4}$

Table 2: Comparison of model selection strategies for languages supported by the MT model. We average the results of TRG, TRG+SRC, M-TRG, and M-TRG+SRC for T-Train, SRC for T-Test, and RT+SRC and M-RT+SRC for RTT. The best results per task and training setup (e.g., T-Train) are underlined; the best results for each training setup are shown in **bold**.

We posit that the choice of high-resource languages critically affects T-Train since the test data is still in the low-resource target language, increasing the risk of negative transfer. In contrast, integrating high-resource languages into RTT (i.e., M-RT-Ens-HR) results in substantial gains of at least 1.8% for AmNLI and NusaX compared to M-RT+SRC. Unlike its success on sequence-level classification tasks, M-RT-Ens-HR degrades performance for Masakha. While ensembles often inherently produce higher scores than single models (Wortsman et al., 2022), our results on sequence-level tasks show that ensembles trained on round-trip translations to various high-resource languages (M-RT-Ens-HR) outperform ensembles trained solely on round-trip translated data to the source language (M-RT-Ens-SRC). In contrast to T-Train, integrating high-resource languages in RTT reduces the likelihood of negative transfer since the test data is in the same language as the training data. Ensembling additionally smooths over language-specific translation and downstream transfer errors. Finally, ensembling monolingual LMs might offer further gains but requires such models for each high-resource language.

MT-Strategies for Model Selection. In XLT, model selection is done using validation data in the source or target language, with the latter violating true ZS-XLT (Schmidt et al., 2022, 2023). The usage of MT to create validation data for model se-

		AmNLI	NusaX	Masakha	Avg
		Zero-Shot			
SRC	X	$44.2_{\pm 0.6}$	$57.8_{\pm 1.4}$	$60.2_{\pm 1.6}$	$54.1_{\pm 1.3}$
		Translate-Train			
TRG+SRC	X	$\mathbf{47.5}_{\pm 0.4}$	$67.5_{\pm 1.3}$	$\mathbf{61.8}_{\pm 0.8}$	$59.0_{\pm 0.9}$
M-TRG+SRC	X	$46.5_{\pm 0.3}$	$\mathbf{74.0}_{\pm 1.4}$	$61.0_{\pm 1.2}$	$\mathbf{60.5}_{\pm 1.1}$
		Translate-Test			
SRC	R	$36.5_{\pm 0.2}$	$54.4_{\pm 1.3}$	$48.1_{\pm 0.5}$	$46.3_{\pm 0.8}$
SRC	X	$37.4_{\pm 0.3}$	$54.9_{\pm 1.5}$	$46.6_{\pm 1.4}$	$46.3_{\pm 1.2}$
		Roundtrip-Train-Test			
M-RT+SRC	R	$38.8_{\pm 0.3}$	$60.5_{\pm 0.7}$	$45.0_{\pm 0.4}$	$48.1_{\pm 0.5}$
M-RT+SRC	X	$39.1_{\pm 0.2}$	$59.1_{\pm 1.2}$	$44.0_{\pm 1.4}$	$47.4_{\pm 1.1}$
		incl. Translations to High-Resource Languages			
M-RT-Ens-HR	X	$41.1_{\pm 0.2}$	$65.0_{\pm 0.6}$	$42.8_{\pm 0.6}$	$49.6_{\pm 0.5}$

Table 3: Results for translation-based XLT for languages **not** supported by the MT model. We use XLM-R (X) and RoBERTa (R). The best results are shown in **bold**.

	AmNLI	NusaX	Masakha	Avg
Nucleus	$56.2_{\pm 3.2}$	$75.6_{\pm 2.4}$	$60.5_{\pm 1.9}$	$64.1_{\pm 2.6}$
Greedy	$62.5_{\pm 0.6}$	$\mathbf{79.5}_{\pm 2.2}$	$64.0_{\pm 1.1}$	$68.7_{\pm 1.5}$
Beam	$\mathbf{62.6}_{\pm 0.5}$	$79.4_{\pm 2.1}$	$\mathbf{64.8}_{\pm 1.2}$	$68.9_{\pm 1.4}$

Table 4: Results for T-Train (TRG) for different decoding strategies evaluated on the validation data of AmNLI, NusaX, and Masakha. The best results are shown in **bold**.

	AmNLI	NusaX	Masakha	Avg
Joint	$63.5_{\pm 0.4}$	$\mathbf{80.7}_{\pm 0.4}$	$\mathbf{62.8}_{\pm 0.4}$	$\mathbf{69.0}_{\pm 0.4}$
Sequential	$\mathbf{64.1}_{\pm 1.4}$	$80.1_{\pm 0.7}$	$62.4_{\pm 0.4}$	$68.9_{\pm 0.9}$

Table 5: Comparison of sequential vs. joint translation-based XLT for languages supported by the MT model. We average the results of TRG+SRC and M-TRG+SRC and the respective sequential variants (SRC→TRG and SRC→M-TRG). The best results are shown in **bold**. Model selection is done on the best epoch based on target language validation data (*Val-Trg*).

lection, however, remains understudied (Ebrahimi et al., 2022). We thus next explore MT-based model selection strategies and compare them against standard counterparts (cf. §3) in Table 2. In T-Train, in line with prior work (Ebrahimi et al., 2022; Schmidt et al., 2022), *Val-Trg* outperforms all other model selection variants. We show, however, for the first time, that it is also the upper bound of T-Test and RTT. Additionally, in T-Train *Val-MT-Trg* (i.e., model selection based on the automatically translated target language validation data) surpasses *Val-Src* on average across all tasks; this is notably not the case for T-Test and RTT.

Unsupported Languages. Even the most multilingual MT models (Team et al., 2022) support only a tiny fraction of the world's 7000+ languages. Table 3 summarizes the performance of our MT-based XLT strategy for languages not supported by MT, where we translate to/from the closest respective supported language (see §2.4). We find that T-Train strategies remain successful and substantially improve by 4.9% (TRG+SRC) and 6.4% (M-TRG+SRC) over the ZS-XLT on average. In contrast, T-Test and RTT for unsupported languages substantially trail ZS-XLT performance. This is because it is not really possible to get good translations in the source language by simply pretending the input comes from a different, supported language (in T-Test and RTT). In contrast, with T-Train, we obtain proper translations in a supported language that is close to the real target (as in T-Train): the transfer then amounts to the mLM-based ZS-XLT ability from the close, MT-supported language to the real MT-unsupported target. This further supports the finding that MT quality much less affects performance of T-Train strategies than of T-Test or RTT approaches (Artetxe et al., 2023).

5 Further Findings

Decoding Strategy. Previous work examined the impact of various decoding strategies on downstream performance, particularly in the context of back-translation (Edunov et al., 2018) and sequence-level classification (Artetxe et al., 2023). They found nucleus sampling consistently superior to beam search and greedy decoding. However, our results in Table 4 suggest a noteworthy deviation for low-resource languages. We find beam search and greedy decoding substantially outperform nucleus sampling. We posit that the underrepresentation of low-resource languages in the training data of MT models contributes to this contrast.[4]

Joint vs. Sequential Training. Prior work primarily concatenated the source data with the translated target language data and trained on both jointly (Hu et al., 2020; Oh et al., 2022; Artetxe et al., 2023). In contrast, Aggarwal et al. (2022) propose a se-

[4]We present details on the resource availability of the tasks we evaluated, compared to related work, in Appendix D.

		AmNLI		NusaX		Masakha		Avg	
		NLLB	GT	NLLB	GT	NLLB	GT	NLLB	GT
Translate-Train									
TRG+SRC	X	62.9	63.1	87.0	87.0	66.0	65.3	72.0	71.8
Translate-Test									
SRC	R	53.1	63.6	85.3	85.7	54.8	59.7	64.4	69.7
SRC	X	52.9	64.8	85.8	86.6	54.1	59.0	64.3	70.1
Roundtrip-Train-Test									
RT+SRC	R	62.4	69.9	85.2	86.9	55.0	59.9	67.5	72.2
RT+SRC	X	63.1	69.1	86.0	87.5	54.0	59.2	67.7	71.9

Table 6: Results for translation-based XLT for two MT systems (NLLB and GT) for languages supported by both MT models. For T-Train, model selection is done on the best epoch based on target-language validation data (*Val-Trg*), and for T-Test and RTT, based on source-language validation data (*Val-Src*). We evaluated XLM-R (X) and RoBERTa (R).

quential T-Train approach in which the model is first trained on the source-language data and then, in a subsequent step, on the translated data of either (i) a single target language (TRG) or (ii) multiple target languages jointly (M-TRG). We adopt this in our T-Train variants (denoted SRC→TRG and SRC→M-TRG) and compare them against the more established joint training: results in Table 5 show comparable performance between the two. This favors sequential training, as it is more computationally efficient (Schmidt et al., 2022).

Importance of the MT model. In the MT landscape, the translation quality of the industrial-grade models is often considered superior to that of their publicly available counterparts. Because of this, we ablate the impact of the used MT model on translation-based XLT by generating translations through Google Translate (GT)—a representative example of an industrial MT model. We evaluate T-Train (i.e., TRG+SRC), T-Test (i.e., SRC), and RTT (i.e., RT+SRC) with GT. Our results in Table 6 indicate that the performance remains comparable for T-Train, while GT surpasses NLLB by a substantial margin in the context of T-Test and RTT. We hypothesize that our observation stems from the increased translation quality which is of larger importance for T-Test and RTT than for T-Train (Artetxe et al., 2023). Furthermore, our ablation confirms that RTT remains the most competitive translation-based XLT method for sequence-level classification tasks. Unfortunately, GT does not support the exact same set of languages as NLLB. We thus carry out the ablation on the following languages supported by both MT systems: for AmNLI: Aymara, Guarani, and Quechua; for NusaX: Javanese and Sundanese; and for Masakha: Bambara, Éwé, Hausa, Igbo, Kinyarwanda, chiShona, Kiswahili, Akan/Twi, isiXhosa, Yorùbá, isiZulu.

6 Related Work

Translation-based Transfer. Translation-based XLT has been adopted early (Fortuna and Shawe-Taylor, 2005; Banea et al., 2008; Shi et al., 2010) yet remains a competitive baseline to date (Ruder et al., 2021; Ebrahimi et al., 2022; Aggarwal et al., 2022). Prior work evaluated training on the translated data of a single target language (Ebrahimi et al., 2022), on the concatenation of all target languages (Ruder et al., 2021), and have integrated the source language either by sequentially training first on the source followed by the translated target language data (Aggarwal et al., 2022) or by jointly training on the concatenation of both (Chen et al., 2023). While earlier approaches focus primarily on the translation of the training data (T-Train), more recent work evaluated the translation of test data as well (Hu et al., 2020; Isbister et al., 2021) (T-Test). Finally, both approaches can be combined by training the model on round-trip translated noisy source data (i.e., translating source data to the target language and back to the source) and evaluating it on target language test data translated to the source language (Artetxe et al., 2020; Oh et al., 2022; Artetxe et al., 2023). Previous studies have either focused on improving one of these paradigms or utilized them as baselines. In contrast, we provide a comparative empirical evaluation of existing translation-based approaches to XLT, testing them explicitly against ZS-XLT for low-resource languages.

Label projection. Translation-based transfer for token-level tasks necessitates label projection, which is achieved through either alignment-based or (Tjong Kim Sang and De Meulder, 2003; Jalili Sabet et al., 2020; Nagata et al., 2020) marker-based approaches (Lee et al., 2018; Lewis et al., 2020; Hu et al., 2020; Bornea et al., 2021). The former maps each token in the source sequence to a token in the translated target sequence, with recent neural word aligners utilizing contextualized embeddings of mLMs to produce the alignment (Dou and Neubig, 2021; Wang et al., 2022). Marker-based alignment, in contrast, entails marking labeled tokens in the sequence prior to translation, often by enclosing them in XML or HTML tags, and preserving them throughout the translation process. Subsequently, the labels can be recovered from the markers. While alignment-based methods are prone to issues like error propagation, translation shift (Akbik et al., 2015), and non-contiguous alignments (Zenkel et al., 2020), marker-based projection compromises translation performance by introducing artificial tokens and is susceptible to vanishing markers, particularly with non-industrial, publicly available translation models (Chen et al., 2023). In XLT for NER (Masakha), we leveraged a state-of-the-art alignment-based model (Wang et al., 2022).

7 Conclusion

We reviewed the field of translation-based cross-lingual transfer (XLT) to low-resource languages through a comparative evaluation of various approaches—derived from translate-train (`T-Train`), translate-test (`T-Test`), and roundtrip-train-test (`RTT`)—on three established benchmarks encompassing 40 languages. We demonstrated that translation-based XLT substantially outperforms zero-shot XLT no matter the task. Furthermore, irrespective of the translation-based strategy, including the clean source language data in the training yielded consistent improvements. For sequence-level tasks, training on the source language data round-trip translated through a set of related target languages and evaluating, at inference, the target language instances translated back to the source language performed best (`RTT`). In contrast, for token-level tasks, training on the translations to a single target language showed the best results (`T-Train`). Additionally, we proposed novel translation-based XLT strategies for `T-Train` and `RTT` by including translations to a set of typologically diverse high-resource languages. Further, we successfully proposed translation-based strategies for languages unsupported by the MT model and showcased the effectiveness of using automatically translated validation data for model selection. Our empirical comparison and its findings warrant broader inclusion of more competitive translation-based XLT approaches as standard baselines in all research efforts set to improve XLT with mLMs.

8 Limitations

We strove to provide a comprehensive and systematic evaluation of translation-based XLT to low-resource languages, additionally providing novel `T-Train` and `RTT` paradigms. However, our study faces limitations, primarily stemming from the prevalent practice of obtaining benchmarks for low-resource languages by translating datasets from high-resource languages, which applies to AmNLI, NusaX, and some languages of Masakha. The resulting data possesses distinctive characteristics arising from the translation process, commonly referred to as *translationese*. On the one hand, we explicitly exploit this behavior by demonstrating that augmenting the training data in the same way as we augment the test data (i.e., `RTT`) yields the best results. On the other hand, there exist uncontrollable implications potentially influencing our results, for instance, that translation often becomes easier for datasets originating from translation themselves.

Acknowledgements

This work was in part supported by the Alcatel-Lucent Stiftung (Grant "EQUIFAIR: Equitably Fair and Trustworthy Language Technology").

References

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Divyanshu Aggarwal, Vivek Gupta, and Anoop Kunchukuttan. 2022. IndicXNLI: Evaluating multilingual inference for Indian languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10994–11006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2015. Generating high quality proposition Banks for multilingual semantic role labeling. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 397–407, Beijing, China. 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.

Ruth-Ann Armstrong, John Hewitt, and Christopher Manning. 2022. JamPatoisNLI: A jamaican patois natural language inference dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5307–5320, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting machine translation for cross-lingual classification.

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

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 127–135, Honolulu, Hawaii. Association for Computational Linguistics.

Mihaela Bornea, Lin Pan, Sara Rosenthal, Radu Florian, and Avirup Sil. 2021. Multilingual transfer learning for qa using translation as data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12583–12591.

Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. 2023. Frustratingly easy label projection for cross-lingual transfer. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5775–5796, Toronto, Canada. Association for Computational Linguistics.

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 ty pologically di verse 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.

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.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, 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.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Blaz Fortuna and John Shawe-Taylor. 2005. The use of machine translation tools for cross-lingual text mining. In *Proceedings of the ICML Workshop on Learning with Multiple Views*. Citeseer.

Goran Glavaš and Ivan Vulić. 2021. Climbing the tower of treebanks: Improving low-resource dependency parsing via hierarchical source selection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4878–4888.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5427–5444, Online. Association for Computational Linguistics.

Michael A. Hedderich, David Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow. 2020. Transfer learning and distant supervision for multilingual transformer models: A study on African languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2580–2591, Online. Association for Computational Linguistics.

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. 2021. Explicit alignment objectives for multilingual bidirectional encoders. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3633–3643, Online. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Tim Isbister, Fredrik Carlsson, and Magnus Sahlgren. 2021. Should we stop training more monolingual models, and simply use machine translation instead? In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 385–390, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

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.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, et al. 2023. Madlad-400: A multilingual and document-level large audited dataset. *arXiv preprint arXiv:2309.04662*.

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.

Kyungjae Lee, Kyoungho Yoon, Sunghyun Park, and Seung-won Hwang. 2018. Semi-supervised training data generation for multilingual question answering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

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

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Sa'id Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alípio Jorge, and Pavel Brazdil. 2022. NaijaSenti: A Nigerian Twitter sentiment corpus for multilingual sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 590–602, Marseille, France. European Language Resources Association.

Masaaki Nagata, Katsuki Chousa, and Masaaki Nishino. 2020. A supervised word alignment method based on cross-language span prediction using multilingual BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 555–565, Online. Association for Computational Linguistics.

Jaehoon Oh, Jongwoo Ko, and Se-Young Yun. 2022. Synergy with translation artifacts for training and inference in multilingual tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6747–6754, Abu Dhabi, United Arab Emirates. 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, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, 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.

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.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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.

Lei Shi, Rada Mihalcea, and Mingjun Tian. 2010. Cross language text classification by model translation and semi-supervised learning. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1057–1067, Cambridge, MA. 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.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

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.

Weikang Wang, Guanhua Chen, Hanqing Wang, Yue Han, and Yun Chen. 2022. Multilingual sentence transformer as a multilingual word aligner. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2952–2963, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zihan Wang, Stephen Mayhew, Dan Roth, et al. 2019. Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840*.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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

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

Shijie Wu and Mark Dredze. 2020. Do explicit alignments robustly improve multilingual encoders? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4471–4482, Online. 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.

Huiyun Yang, Huadong Chen, Hao Zhou, and Lei Li. 2022. Enhancing cross-lingual transfer by manifold mixup. In *International Conference on Learning Representations*.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2020. End-to-end neural word alignment outperforms GIZA++. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1605–1617, Online. Association for Computational Linguistics.

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

A Models and Datasets

Language	Code	Pre. Model	Supp. Trans.	Closest
AmNLI				
Aymara	AYM	No	Yes	-
Guaraní	GN	No	Yes	-
Quechua	QUY	No	Yes	-
Asháninka	CNI	No	No	AYM
Bribri	BZD	No	No	QUY
Nahuatl	NAH	No	No	GN
Otomí	OTO	No	No	GN
Rarámuri	TAR	No	No	AYM
Shipibo-Konibo	SHP	No	No	QUY
Wixarika	HCH	No	No	GN
NusaX				
Acehnese	ACE	No	Yes	-
Balinese	BAN	No	Yes	-
Banjarese	BJN	No	Yes	-
Buginese	BUG	No	Yes	-
Minangkabau	MIN	No	Yes	-
Javanese	JAV	Yes	Yes	-
Sundanese	SUN	Yes	Yes	-
Madurese	MAD	No	No	SUN
Ngaju	NIJ	No	No	SUN
Toba Batak	BBC	No	No	BUG
MasakhaNER				
Bambara	BAM	No	Yes	-
Éwé	EWE	No	Yes	-
Fon	FON	No	Yes	-
Hausa	HAU	Yes	Yes	-
Igbo	IBO	No	Yes	-
Kinyarwanda	KIN	No	Yes	-
Luganda	LUG	No	Yes	-
Luo	LUO	No	Yes	-
Mossi	MOS	No	Yes	-
Chichewa	NYA	No	Yes	-
chiShona	SNA	No	Yes	-
Kiswahili	SWA	Yes	Yes	-
Setswana	TSN	No	Yes	-
Akan/Twi	TWI	No	Yes	-
Wolof	WOL	No	Yes	-
isiXhosa	XHO	Yes	Yes	-
Yorùbá	YOR	No	Yes	-
isiZulu	ZUL	No	Yes	-
Ghomálá'	BBJ	No	No	SWA
Naija	PCM	No	No	HAU

Table 7: List of languages per task showing the coverage in the pretraining corpus of our backbone model (*Pre. Model*), the support by NLLB (*Supp. Trans.*), and the closest language we translated to/from for languages that are not supported by the translation model (*Closest*).

The models for translation, word alignment, and downstream fine-tuning were accessed through the Hugging Face transformers library (Wolf et al., 2020). Additional adapter checkpoints for the used word aligner were downloaded from the corresponding GitHub repository: AccAlign (Wang et al., 2022). We accessed all our datasets through the Hugging Face datasets library (Lhoest et al.,

	AccAlign	EasyProj
BAM	94.4	90.9
EWE	95.6	92.2
FON	92.9	83.4
HAU	97.5	94.4
IBO	98.3	96.1
KIN	97.2	93.8
LUG	97.0	95.3
LUO	96.6	94.0
MOS	90.3	92.3
NYA	98.5	96.7
SNA	98.7	95.6
SWA	98.8	96.3
TSN	98.0	95.0
TWI	96.2	94.6
WOL	93.0	93.4
XHO	97.8	95.1
YOR	97.3	94.3
ZUL	97.7	93.1
Avg Proj. Rate	96.4	93.7
Avg. Perf.	65.5	65.6

Table 8: Projection rates and average performance in the TRG+SRC setup for word alignments produce by AccAlign (Wang et al., 2022) and EasyProj (Chen et al., 2023). Model selection is done on the best epoch based on target-language validation data (*Val-Trg*).

2021). Further, we ensured compliance with the licenses of the models and datasets. Table 7 displays a detailed list of all languages.

B Word Alignment

Table 8 shows the projection rates for AccAlign (Wang et al., 2022) (used in our work) and the state-of-the-art marker-based method EasyProject (EasyProj) (Chen et al., 2023). The projection rate is computed as the ratio of retained training instances after label projection to all instances in the original training data. The results highlight that the downstream performance of AccAlign is on par with the competitive EasyProj. Nevertheless, we attribute variations in the projection rate not only to superior alignment but also to differences in filtering strategies. While Chen et al. (2023) filter translated instances that do not match the number and type of tags in the source instance, our approach filters instances if a tagged source-language token cannot be mapped to its target language equivalent. We leave the exploration of the impact of different filtering approaches to future work.

	AmNLI	NusaX	Masakha
Task	NLI	TC	NER
Epochs	2	20	10
Batch Size	32	32	32
Learning Rate	2e-6	1e-5	1e-5
Weight Decay	0.01	0.01	0.01

Table 9: Hyperparameters for downstream fine-tuning.

C Training and Computational Details

Table 9 outlines the hyperparameters for downstream fine-tuning of our utilized tasks.[5] Alongside, we implement a linear schedule of 10% warm-up and decay and employ mixed precision. All translations were run on a single A100 with 40GB VRAM, and all downstream training and evaluation runs were completed on a single V100 with 32GB VRAM. We roughly estimate that GPU time accumulates to 3500 hours across all translations and downstream fine-tunings.

D Resource Availability

To substantiate our claim that the languages we evaluate are characterized by far lower resource availability compared to related work, we assess the relative size of parallel data used for training NLLB for languages encompassed in the datasets we used and those employed in Artetxe et al. (2023). For each language, we calculate the ratio of available parallel data to the total size of the parallel corpus and, subsequently, average the results per dataset. The computations are based on the following corpus https://huggingface.co/datasets/allenai/nllb. Our metric serves as a proxy for the average coverage of a dataset in the training data of NLLB. As shown in Table 10, the resource availability for the datasets we evaluated is approximately an order of magnitude smaller.

[5] We used a comparably small learning rate for AmNLI as single seeds did not converge for higher learning rates in preliminary experiments.

	Artetxe et al. (2023)							Ours			
	XNLI	PAWS-X	MARC	XStory	XCOPA	EXAMS	Avg	AmNLI	NusaX	Masakha	Avg
Avg. Res. Availability	2.67	3.61	4.24	2.24	1.17	2.67	2.78	0.09	0.21	0.41	0.24

Table 10: Average percentage of available parallel data per task from the corpus used to train NLLB for three datasets we evaluated on: AmNLI, NusaX, and Masakha; and six datasets Artetxe et al. (2023) did: XNLI (Conneau et al., 2018), PAWS-X (Yang et al., 2019), MARC (Keung et al., 2020), XStoryCloze (XStory) (Lin et al., 2022), XCOPA (Ponti et al., 2020), EXAMS (Hardalov et al., 2020).

E Detailed Main Results

		AYM			GN			QUY			Avg		
		I	II	III	I	II	III	I	II	III	I	II	III
Zero-Shot													
SRC	X	43.2	44.0	42.4	46.5	46.8	47.7	44.3	44.7	44.2	44.7	45.2	44.8
Translate-Train													
SRC+HR	X	38.0	38.8	38.8	42.0	44.8	44.5	40.2	42.1	41.7	40.1	41.9	41.7
T	X	58.4	59.5	58.7	63.6	63.2	62.8	61.5	62.2	61.8	61.1	61.6	61.1
TRG+SRC	X	59.4	59.4	59.6	66.1	65.6	65.6	61.9	62.3	63.4	62.4	62.4	62.9
SRC→TRG	X	53.8	62.8	61.9	64.1	66.2	67.0	54.8	64.3	62.7	57.6	64.4	63.9
TRG+SRC+HR	X	59.4	59.7	59.8	65.8	65.8	66.2	63.5	63.9	64.3	62.9	63.1	63.4
M-TRG	X	61.2	61.6	61.4	64.4	64.1	64.2	64.7	64.4	64.7	63.4	63.4	63.5
M-TRG+SRC	X	61.4	62.4	62.3	65.5	65.2	65.2	63.8	64.0	64.8	63.6	63.9	64.1
SRC→M-TRG	X	58.3	62.1	62.6	60.6	66.8	66.8	59.5	65.0	65.0	59.5	64.7	64.8
M-TRG+SRC+HR	X	62.7	63.0	62.7	66.6	67.0	66.3	64.7	64.6	65.1	64.7	64.9	64.7
Translate-Test													
SRC	R	46.9	46.9	46.9	60.2	60.1	60.0	52.3	52.5	52.8	53.1	53.2	53.2
SRC	X	46.3	46.3	47.8	60.8	61.0	60.8	51.7	52.0	52.5	52.9	53.1	53.7
Roundtrip-Train-Test													
RT+SRC	R	58.1	59.2	58.4	68.5	67.6	68.2	60.6	61.3	61.3	62.4	62.7	62.6
RT+SRC	X	58.9	59.3	59.3	69.7	69.7	69.3	60.7	60.4	60.6	63.1	63.1	63.1
M-RT+SRC	R	60.8	61.0	60.4	69.6	69.0	69.2	62.4	62.6	62.0	64.3	64.2	63.9
M-RT+SRC	X	59.8	59.7	59.6	69.6	69.5	69.3	62.7	62.9	62.9	64.0	64.0	63.9
M-RT-Ens-SRC	X	59.6	60.0	60.1	70.1	69.9	69.2	61.4	62.3	62.7	63.7	64.0	64.0
M-RT-Ens-HR	X	61.1	61.6	62.7	70.3	70.1	70.0	66.8	66.1	66.1	66.1	65.9	66.3

Table 11: Results for translation-based XLT evaluated of AmNLI for languages supported by the translation model. Model selection is done on the best epoch based on source-language validation data (*Val-Src* (I)), based on translated source-language validation data (*Val-MT-Trg* (II)), and based on target-language validation data (*Val-Trg* (III)). We use XLM-R (X) and RoBERTa (R).

		BZD			CNI			HCH			NAH			OTO			SHP			TAR			AVG		
		I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III
Zero-Shot																									
SRC	X	44.1	42.4	44.5	44.0	44.1	44.9	40.9	40.9	40.7	45.9	45.9	46.5	43.8	44.0	44.1	50.5	50.0	49.7	40.1	43.4	44.4	44.2	44.4	45.0
Translate-Train																									
SRC+HR	X	42.0	42.0	43.6	40.9	43.9	43.9	36.1	39.2	38.1	43.1	44.2	44.0	43.8	44.0	44.1	45.1	48.5	46.8	38.2	41.5	42.5	41.3	43.3	43.3
TRG	X	43.2	42.6	45.0	48.8	46.4	48.4	44.6	46.1	46.4	49.3	49.2	49.5	47.5	47.4	46.8	50.5	49.1	50.7	47.7	49.2	49.1	47.4	47.1	48.0
TRG+SRC	X	44.9	44.4	45.7	47.6	47.5	48.8	44.8	45.0	45.7	48.4	48.4	48.6	47.8	47.8	48.0	51.0	48.0	51.0	47.7	48.5	49.0	47.5	47.1	48.1
SRC→TRG	X	46.1	44.2	45.7	47.8	48.0	48.9	45.7	46.0	45.4	47.9	47.4	49.3	47.2	48.9	47.6	49.7	49.7	49.6	45.4	46.5	47.1	47.1	47.2	47.7
TRG+SRC+HR	X	44.5	44.4	44.9	46.8	47.0	47.6	44.7	44.8	45.6	49.2	50.1	48.9	47.3	48.1	47.4	48.1	47.8	49.0	49.4	49.1	49.6	47.2	47.3	47.6
M-TRG	X	45.9	44.9	46.2	46.1	46.1	45.6	45.0	44.8	45.1	49.6	49.1	48.6	46.3	46.9	45.5	48.5	48.9	48.8	47.2	46.6	49.5	46.9	46.8	47.0
M-TRG+SRC	X	45.5	45.5	46.1	45.5	46.6	46.7	44.4	44.9	44.6	48.1	47.7	48.7	46.9	46.9	46.1	49.5	49.2	50.2	45.6	45.8	46.4	46.5	46.7	47.0
SRC→M-TRG	X	46.4	46.0	45.5	47.4	47.4	47.0	45.7	45.3	45.1	48.5	47.4	48.8	47.8	47.5	47.6	50.8	49.2	51.0	46.8	46.0	47.0	47.6	47.0	47.4
M-TRG+SRC+HR	X	45.2	45.3	46.9	45.8	46.5	46.5	45.2	45.1	45.0	48.2	48.6	50.1	47.0	47.4	47.1	50.0	50.1	50.7	46.3	46.2	47.7	46.8	47.0	47.7
Translate-Test																									
SRC	R	35.8	36.0	35.6	32.9	33.7	33.1	36.5	36.1	36.9	39.5	40.1	39.6	38.4	38.2	37.3	38.5	38.8	39.2	33.8	34.4	33.5	36.5	36.8	36.5
SRC	X	35.3	35.1	36.1	35.8	36.4	36.4	37.0	37.1	36.3	38.8	39.2	38.7	39.4	39.4	38.0	41.4	40.9	40.8	33.8	34.3	33.9	37.4	37.5	37.2
Roundtrip-Train-Test																									
RT+SRC	R	36.4	36.7	36.8	36.5	36.9	36.7	37.3	36.5	37.2	39.8	39.8	39.5	41.5	40.6	40.6	42.7	42.1	41.5	34.5	34.7	34.0	38.4	38.2	38.0
RT+SRC	X	37.4	36.2	35.8	37.4	37.2	36.7	37.3	37.3	36.8	39.5	39.6	39.2	40.4	40.2	40.9	43.5	44.4	43.2	35.1	35.8	35.0	38.6	38.7	38.2
M-RT+SRC	R	37.1	37.5	37.1	38.9	39.3	37.9	38.4	37.9	38.6	39.4	39.4	40.2	40.9	40.8	41.8	41.9	41.7	43.3	35.3	34.7	34.5	38.8	38.7	39.0
M-RT+SRC	X	37.1	36.8	37.2	39.0	38.8	37.8	39.6	39.4	39.5	41.1	40.4	41.0	39.3	39.8	39.4	43.2	42.8	42.9	34.8	34.8	34.9	39.1	39.0	38.9
M-RT-Ens-SRC	X	37.0	37.0	36.8	38.7	39.0	38.2	39.2	38.6	39.4	41.3	40.2	40.3	39.4	38.6	41.2	43.5	42.5	43.2	34.8	34.5	35.1	39.1	38.6	39.2
M-RT-Ens-HR	X	41.1	40.7	41.4	39.1	38.9	39.2	39.9	40.7	40.5	43.7	43.3	44.9	40.2	40.9	42.2	46.6	47.2	46.7	37.4	38.3	37.6	41.1	41.4	41.8

Table 12: Results for translation-based XLT evaluated of AmNLI for languages **not** supported by the translation model. Model selection is done on the best epoch based on source-language validation data (*Val-Src* (I)), based on translated source-language validation data (*Val-MT-Trg* (II)), and based on target-language validation data (*Val-Trg* (III)). We use XLM-R (X) and RoBERTa (R).

		ACE			BAN			BJN			BUG			JAV			MIN			SUN			Avg		
		I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III
Zero-Shot																									
SRC	X	65.7	64.6	65.7	72.5	72.7	71.9	79.5	79.7	80.1	36.9	42.6	43.9	82.7	79.9	84.8	79.2	80.3	80.4	81.8	83.9	83.6	71.2	72.0	72.9
Translate-Train																									
SRC+HR	X	67.0	68.0	68.8	72.0	72.5	73.0	80.4	80.4	80.6	39.6	44.1	43.5	80.7	83.7	86.0	77.2	79.1	78.9	81.3	80.8	81.0	71.2	72.7	73.1
TRG	X	74.1	74.4	75.3	73.2	75.5	74.0	83.4	82.7	82.1	62.2	64.6	64.7	86.1	86.0	88.9	82.2	83.2	83.1	83.6	83.4	84.3	77.8	78.6	78.9
TRG+SRC	X	76.2	75.6	77.6	76.8	75.9	75.6	82.4	83.4	82.0	65.1	65.6	67.0	88.1	87.1	90.9	85.1	84.6	85.3	84.6	84.1	83.0	79.7	79.5	80.2
SRC→TRG	X	74.6	75.0	75.6	76.3	77.0	77.0	81.9	82.1	82.6	64.7	62.9	65.4	86.9	87.2	89.7	83.5	84.3	82.9	84.1	83.9	83.6	78.9	78.9	79.5
TRG+SRC+HR	X	73.1	75.1	75.4	75.4	76.2	76.1	81.5	81.6	82.2	63.6	61.8	64.6	87.4	87.8	89.5	82.6	83.8	84.7	83.1	84.4	83.9	78.1	78.7	79.5
M-TRG	X	74.8	77.8	77.9	75.6	77.5	77.1	84.1	84.3	84.5	65.0	64.6	65.2	84.8	86.0	88.8	85.1	84.0	84.3	83.6	84.4	84.6	79.0	79.8	80.3
M-TRG+SRC	X	77.7	77.8	76.4	77.4	77.3	78.5	86.1	84.8	86.0	65.1	66.2	66.8	86.5	84.4	88.3	86.5	85.8	86.2	86.5	86.4	86.5	80.8	80.4	81.2
SRC→M-TRG	X	75.3	76.7	75.4	77.1	78.5	78.0	84.2	83.6	85.0	64.0	66.8	67.7	84.0	83.2	88.2	84.3	84.4	84.4	85.6	85.7	83.5	79.2	79.8	80.3
M-TRG+SRC+HR	X	76.8	78.2	77.8	76.5	78.0	77.1	84.0	84.1	84.9	65.6	66.9	66.0	81.8	84.9	88.5	83.5	83.9	85.1	85.5	85.1	85.4	79.1	80.2	80.7
Translate-Test																									
SRC	R	77.3	75.5	77.5	74.1	75.5	75.8	82.2	79.6	82.0	69.5	71.8	72.3	85.8	84.3	85.5	81.9	82.0	82.9	84.8	84.3	84.8	79.4	79.0	80.1
SRC	X	78.8	77.9	78.5	77.2	77.4	78.8	83.6	83.3	82.3	71.7	70.1	74.5	85.5	86.1	85.8	83.4	83.6	84.6	86.1	86.3	85.8	80.9	80.7	81.5
Roundtrip-Train-Test																									
RT+SRC	R	79.5	79.3	79.1	76.1	77.9	77.8	82.8	82.4	82.2	74.5	74.3	73.7	85.7	83.7	85.0	85.6	84.3	84.3	84.6	84.7	85.3	81.2	81.0	81.0
RT+SRC	X	78.3	79.4	79.7	78.8	78.1	77.1	83.9	83.8	84.1	73.1	74.1	75.3	86.5	86.8	86.4	84.9	85.5	85.9	85.6	84.9	85.7	81.6	81.8	82.0
M-RT+SRC	R	78.6	77.8	78.2	77.8	79.3	80.1	83.6	83.6	83.4	73.8	73.4	74.6	85.8	85.3	86.0	83.9	84.2	84.2	83.7	83.6	83.5	81.0	81.0	81.4
M-RT+SRC	X	78.8	78.6	79.8	79.6	78.0	80.3	85.2	84.8	85.0	74.5	75.1	75.6	86.9	87.1	86.6	84.7	84.3	84.8	85.0	85.1	84.8	82.1	81.9	82.4
M-RT-Ens-SRC	X	79.8	79.1	80.2	80.2	80.0	80.5	86.5	86.5	86.3	74.8	75.8	75.7	87.5	87.3	86.6	85.3	85.8	85.3	85.8	85.7	84.2	82.8	82.9	82.7
M-RT-Ens-HR	X	83.2	83.5	83.2	82.2	81.6	82.4	86.0	85.7	85.1	75.2	75.5	74.6	88.0	88.0	87.0	86.5	86.1	86.7	86.5	86.9	86.0	83.9	83.9	83.6

Table 13: Results for translation-based XLT evaluated of NusaX for languages supported by the translation model. Model selection is done on the best epoch based on source-language validation data (*Val-Src* (I)), based on translated source-language validation data (*Val-MT-Trg* (II)), and based on target-language validation data (*Val-Trg* (III)). We use XLM-R (X) and RoBERTa (R).

		BBC			MAD			NIJ			Avg		
		I	II	III	I	II	III	I	II	III	I	II	III
Zero-Shot													
SRC	X	41.4	45.5	45.9	65.5	64.8	67.4	66.6	65.7	67.2	57.8	58.7	60.2
Translate-Train													
SRC+HR	X	42.7	46.5	45.3	65.1	62.8	68.5	62.7	62.1	65.6	56.8	57.1	59.8
TRG	X	60.6	60.9	62.2	70.1	69.6	73.1	66.1	67.4	69.9	65.6	65.9	68.4
TRG+SRC	X	61.2	62.2	64.0	72.4	72.4	71.9	69.0	69.7	68.6	67.5	68.1	68.2
SRC→TRG	X	63.8	62.7	66.1	70.7	69.5	70.7	69.6	69.2	70.1	68.0	67.1	68.9
TRG+SRC+HR	X	62.2	63.2	61.7	71.7	72.1	72.4	71.5	68.9	71.6	68.5	68.1	68.6
M-TRG	X	65.8	67.8	66.8	76.8	75.6	78.9	74.8	74.5	76.2	72.5	72.6	74.0
M-TRG+SRC	X	66.3	67.9	65.2	78.2	76.6	77.8	77.5	75.7	77.8	74.0	73.4	73.6
SRC→M-TRG	X	68.0	68.3	65.6	77.8	77.9	78.0	76.1	77.2	78.4	74.0	74.5	74.0
M-TRG+SRC+HR	X	65.1	66.9	64.0	76.7	75.5	77.0	75.2	75.3	76.8	72.3	72.6	72.6
Translate-Test													
SRC	R	42.6	47.8	49.2	56.4	56.4	58.8	64.3	63.7	65.8	54.4	56.0	57.9
SRC	X	40.4	38.5	55.1	60.9	59.8	63.6	63.4	62.1	65.8	54.9	53.5	61.5
Roundtrip-Train-Test													
RT+SRC	R	49.6	45.5	50.2	55.1	56.1	58.1	60.9	62.0	63.1	55.2	54.5	57.1
RT+SRC	X	44.0	46.6	54.6	62.5	62.2	64.3	64.6	65.2	64.3	57.0	58.0	61.1
M-RT+SRC	R	51.5	50.5	52.5	61.7	60.8	61.6	68.3	66.4	68.3	60.5	59.2	60.8
M-RT+SRC	X	47.2	54.2	55.3	62.7	64.6	66.7	67.3	68.5	67.0	59.1	62.4	63.0
M-RT-Ens-SRC	X	49.4	54.1	56.4	65.7	68.0	68.5	68.3	69.5	69.6	61.1	63.9	64.8
M-RT-Ens-HR	X	51.9	56.9	58.1	69.8	68.8	70.9	73.2	72.5	72.7	65.0	66.1	67.2

Table 14: Results for translation-based XLT evaluated of NusaX for languages **not** supported by the translation model. Model selection is done on the best epoch based on source-language validation data (*Val-Src* (I)), based on translated source-language validation data (*Val-MT-Trg* (II)), and based on target-language validation data (*Val-Trg* (III)). We use XLM-R (X) and RoBERTa (R).

		BAM	EWE	FON	HAU	IBO	KIN	LUG	LUO	MOS	NYA	SNA	SWA	TSN	TWI	WOL	XHO	YOR	ZUL	Avg
Zero-Shot																				
SRC	X	36.9	67.9	46.8	73.4	48.0	42.0	58.6	37.7	47.6	47.4	35.8	85.5	48.1	43.3	48.2	22.8	31.1	41.1	47.9
Translate-Train																				
SRC+HR	X	35.8	70.4	50.5	72.5	54.6	43.3	63.9	40.9	50.6	53.3	55.4	81.9	52.5	42.7	52.3	56.9	33.6	59.1	53.9
TRG	X	51.3	72.6	64.5	71.4	65.8	53.5	69.7	47.7	53.9	63.9	65.6	76.3	68.0	60.3	58.8	68.7	36.4	69.0	62.1
TRG+SRC	X	48.9	75.4	65.9	72.3	68.1	54.7	74.3	50.1	57.0	68.0	69.9	77.4	68.7	61.5	61.7	70.0	38.0	71.4	64.1
SRC→TRG	X	51.0	71.4	65.7	72.1	68.3	54.2	73.2	48.8	56.1	65.0	67.4	76.2	69.7	60.1	56.9	69.1	38.1	70.9	63.0
TRG+SRC+HR	X	47.9	71.8	67.2	71.5	70.2	54.4	73.3	48.6	54.5	66.6	68.1	76.1	68.0	61.0	58.0	68.5	38.0	68.7	62.9
M-TRG	X	43.8	65.1	60.7	69.2	63.7	51.1	66.2	47.1	45.2	57.0	62.1	75.2	58.2	58.9	48.4	58.1	36.2	58.6	56.9
M-TRG+SRC	X	44.1	65.5	58.1	70.1	61.8	53.2	66.7	45.6	46.7	56.6	60.7	76.1	62.4	59.7	46.8	61.2	35.9	62.8	57.4
SRC→M-TRG	X	48.3	68.0	63.7	69.8	64.8	54.0	67.0	48.4	50.1	58.6	61.2	75.9	61.2	60.1	52.5	62.0	38.7	62.5	59.3
M-TRG+SRC+HR	X	45.7	65.4	64.3	69.0	64.9	52.7	65.2	46.5	49.2	56.9	60.7	75.3	57.9	58.6	53.9	60.4	36.7	61.7	58.0
Translate-Test																				
SRC	R	39.9	61.3	56.4	58.0	55.8	51.6	68.1	45.5	39.6	63.7	58.5	62.0	60.1	56.8	49.7	58.0	43.9	57.0	54.8
SRC	X	39.4	61.5	56.3	57.8	54.9	50.5	67.9	43.2	39.1	63.1	58.0	61.6	57.9	55.2	49.9	57.6	43.4	56.7	54.1
Roundtrip-Train-Test																				
RT+SRC	R	39.7	61.2	57.2	58.3	60.6	49.9	65.6	44.0	37.6	63.8	57.8	62.2	59.8	57.2	50.9	55.9	45.2	57.4	54.7
RT+SRC	X	39.0	60.0	57.2	57.8	58.2	50.6	65.1	42.6	36.4	62.5	57.0	61.8	57.6	55.2	50.1	55.0	44.3	56.5	53.7
M-RT+SRC	R	40.0	57.9	55.0	58.3	59.8	49.6	63.7	43.0	35.5	62.3	55.3	62.7	59.5	55.6	50.1	54.5	43.7	55.4	53.4
M-RT+SRC	X	39.1	59.0	55.8	58.4	59.6	49.3	65.1	41.1	36.9	61.1	55.3	62.4	58.2	56.0	50.2	53.5	43.0	55.7	53.3
M-RT-Ens-SRC	X	39.9	59.2	56.2	58.0	60.3	50.2	64.8	41.5	38.1	61.7	56.0	62.4	57.0	56.5	50.9	54.1	44.2	55.8	53.7
M-RT-Ens-HR	X	33.8	50.6	46.7	47.6	50.4	42.1	53.7	34.7	34.7	53.1	50.2	54.1	48.3	47.0	44.3	47.9	38.6	46.6	45.8

Table 15: Results for translation-based XLT evaluated of Masakha for languages supported by the translation model. Model selection is done on the best epoch based on source-language validation data (*Val-Src*). We use XLM-R (X) and RoBERTa (R).

		BAM	EWE	FON	HAU	IBO	KIN	LUG	LUO	MOS	NYA	SNA	SWA	TSN	TWI	WOL	XHO	YOR	ZUL	Avg
										Zero-Shot										
SRC	X	38.9	69.1	49.4	73.2	50.6	43.3	62.4	38.4	49.8	49.0	35.7	85.3	49.6	45.2	50.9	22.6	32.4	41.3	49.3
										Translate-Train										
SRC+HR	X	38.7	72.4	54.4	72.6	58.5	46.0	65.5	40.6	51.9	54.4	54.6	82.0	52.9	47.7	51.6	57.3	33.7	59.4	55.2
TRG	X	50.0	74.4	65.0	71.2	65.7	53.8	73.0	48.4	55.0	64.6	66.3	76.4	68.6	58.7	58.8	68.2	37.1	70.2	62.5
TRG+SRC	X	50.6	75.5	66.0	72.2	68.6	55.3	75.0	50.0	55.6	67.5	69.2	77.7	69.5	61.8	60.6	69.3	38.6	72.1	64.2
SRC→TRG	X	50.8	73.6	66.0	72.0	68.5	56.0	74.9	50.0	56.1	67.0	70.0	76.9	69.9	61.8	61.0	69.5	38.4	71.3	64.1
TRG+SRC+HR	X	50.9	72.0	67.5	71.6	69.7	54.1	73.9	49.1	54.1	67.5	69.3	76.8	68.7	61.5	57.9	68.8	39.1	70.2	63.5
M-TRG	X	46.5	64.4	59.0	69.4	64.1	51.4	65.4	45.3	46.7	56.9	60.3	74.6	59.6	58.3	52.4	59.6	36.6	60.0	57.2
M-TRG+SRC	X	44.7	65.5	59.4	68.8	66.2	51.7	63.3	46.8	47.4	56.7	60.3	74.8	59.9	57.4	53.5	59.8	36.0	61.8	57.4
SRC→M-TRG	X	45.1	64.6	62.9	69.6	65.1	51.6	67.3	46.1	46.4	55.6	60.4	73.8	59.1	60.1	51.1	61.3	34.9	61.1	57.6
M-TRG+SRC+HR	X	45.0	64.0	59.5	68.4	65.5	51.0	62.3	47.1	48.0	56.5	60.3	74.6	59.1	57.9	51.7	59.3	33.6	61.0	56.9
										Translate-Test										
SRC	R	39.9	61.4	56.3	58.0	55.9	51.4	67.8	44.8	39.7	63.7	58.5	62.0	60.0	56.3	49.8	57.6	44.1	57.4	54.7
SRC	X	39.2	61.2	55.6	57.8	54.8	50.6	67.7	43.0	39.1	63.3	57.8	61.7	57.8	54.5	49.7	57.5	43.0	56.6	53.9
										Roundtrip-Train-Test										
RT+SRC	R	39.7	61.1	56.9	58.4	61.2	50.2	65.9	44.5	37.7	63.5	57.8	62.4	59.8	56.9	51.4	55.9	45.5	57.0	54.8
RT+SRC	X	39.7	59.6	56.5	58.0	58.6	50.7	66.0	42.5	36.4	62.7	57.1	61.9	57.0	54.1	50.7	55.1	44.3	56.0	53.7
M-RT+SRC	R	39.1	58.4	56.6	58.1	60.2	49.1	62.3	42.0	35.1	62.3	56.2	62.2	59.8	57.2	50.3	54.5	43.2	55.7	53.5
M-RT+SRC	X	40.4	57.2	55.4	58.1	61.2	49.0	62.5	42.0	35.9	61.5	54.7	62.1	57.8	55.7	49.7	50.9	42.6	54.2	52.8
M-RT-Ens-SRC	X	40.2	58.8	55.9	58.3	60.7	50.0	64.7	40.8	36.8	61.8	56.0	62.7	57.1	56.2	51.2	53.7	43.7	55.8	53.6
M-RT-Ens-HR	X	33.9	50.6	47.1	47.9	50.3	41.9	53.3	34.2	34.5	53.3	49.8	54.7	48.3	47.4	44.5	47.8	37.8	46.9	45.8

Table 16: Results for translation-based XLT evaluated of Masakha for languages supported by the translation model. Model selection is done on the best epoch based on translated source-language validation data (*Val-MT-Trg*). We use XLM-R (X) and RoBERTa (R).

		BAM	EWE	FON	HAU	IBO	KIN	LUG	LUO	MOS	NYA	SNA	SWA	TSN	TWI	WOL	XHO	YOR	ZUL	Avg
										Zero-Shot										
SRC	X	39.6	70.8	50.9	73.4	52.9	43.5	64.7	39.3	49.6	51.6	40.6	85.5	52.7	46.0	51.6	22.2	34.4	41.9	50.6
										Translate-Train										
SRC+HR	X	40.3	72.3	56.2	72.9	60.9	46.4	66.0	39.9	53.9	54.1	55.9	84.0	53.4	49.5	53.8	57.2	35.2	60.9	56.3
TRG	X	52.0	75.5	64.7	71.3	66.8	54.5	75.0	49.5	59.3	64.6	67.9	76.6	67.8	61.8	59.5	67.9	37.7	69.8	63.4
TRG+SRC	X	54.6	77.1	67.1	72.6	69.9	56.8	76.5	50.9	58.5	68.3	70.2	79.2	69.8	62.4	61.8	70.1	40.2	72.9	65.5
SRC→TRG	X	52.0	75.5	66.8	72.8	69.5	56.8	76.5	49.3	59.1	68.0	70.1	77.9	69.8	61.3	61.6	69.9	39.7	71.7	64.9
TRG+SRC+HR	X	52.9	74.4	68.1	73.0	70.2	55.0	74.7	49.1	57.8	69.1	70.0	77.8	68.5	61.9	60.2	69.5	40.1	69.6	64.5
M-TRG	X	49.1	71.7	63.4	71.3	66.2	54.9	66.2	47.6	49.3	58.4	63.0	76.9	62.9	57.7	54.9	63.4	37.8	63.7	59.9
M-TRG+SRC	X	49.0	70.0	61.2	71.0	67.3	53.5	69.4	47.1	50.4	59.1	62.7	76.8	62.3	58.3	55.9	62.6	39.1	64.7	60.0
SRC→M-TRG	X	48.7	70.3	64.7	70.8	66.6	55.1	69.6	49.4	50.4	59.9	62.2	76.0	62.1	59.2	52.6	63.0	38.2	63.7	60.1
M-TRG+SRC+HR	X	48.6	68.3	64.5	70.9	68.2	53.9	68.9	45.8	49.3	59.3	62.9	76.6	63.3	61.8	55.5	63.0	39.2	63.1	60.2
										Translate-Test										
SRC	R	39.7	61.4	56.6	58.0	56.5	51.6	67.9	45.0	39.7	63.6	58.4	61.8	59.6	56.9	49.8	57.8	44.3	57.0	54.7
SRC	X	39.8	61.1	56.1	57.8	55.1	50.7	67.9	42.5	38.9	63.3	57.7	61.6	57.8	54.9	49.6	57.3	43.0	56.7	54.0
										Roundtrip-Train-Test										
RT+SRC	R	40.6	60.3	56.5	58.3	61.1	51.1	66.9	43.4	38.0	63.6	57.8	62.6	60.3	57.0	51.7	55.7	45.4	56.0	54.8
RT+SRC	X	40.4	60.4	57.3	58.2	58.3	50.6	66.9	41.8	37.1	63.0	56.9	62.4	55.7	56.8	51.0	55.2	44.6	56.9	54.1
M-RT+SRC	R	40.9	59.1	55.9	57.9	61.6	50.1	65.1	42.7	36.5	62.4	55.5	63.7	59.6	57.2	50.7	53.7	44.1	55.7	54.0
M-RT+SRC	X	40.7	59.3	55.4	58.1	61.5	49.4	63.9	40.2	36.6	61.2	55.0	63.1	58.2	56.3	49.9	50.0	44.0	54.9	53.2
M-RT-Ens-SRC	X	40.5	59.8	55.9	58.2	61.0	50.6	65.5	41.7	38.1	62.0	56.1	63.3	57.7	57.3	51.0	51.7	44.7	55.2	53.9
M-RT-Ens-HR	X	35.4	52.4	48.3	48.4	51.8	43.3	57.1	35.4	35.5	53.4	50.6	55.3	49.3	49.1	45.4	48.1	40.2	49.4	47.1

Table 17: Results for translation-based XLT evaluated of Masakha for languages supported by the translation model. Model selection is done on the best epoch based on target-language validation data (*Val-Trg*). We use XLM-R (X) and RoBERTa (R).

		BBJ			PCM			Avg		
		I	II	III	I	II	III	I	II	III
Zero-Shot										
SRC	X	41.9	42.0	45.4	78.5	78.3	78.2	60.2	60.1	61.8
Translate-Train										
SRC+HR	X	45.8	45.7	44.6	77.2	77.3	76.5	61.5	61.5	60.6
TRG	X	43.2	41.8	44.1	75.0	75.7	75.9	59.1	58.7	60.0
TRG+SRC	X	46.3	46.5	48.7	77.3	77.2	77.6	61.8	61.8	63.2
SRC→TRG	X	46.0	46.7	47.5	76.3	76.4	77.1	61.2	61.5	62.3
TRG+SRC+HR	X	42.0	44.8	46.3	77.0	77.0	77.4	59.5	60.9	61.9
M-TRG	X	48.4	48.1	49.9	72.6	73.7	73.0	60.5	60.9	61.4
M-TRG+SRC	X	47.9	47.0	51.0	74.0	72.9	75.8	61.0	60.0	63.4
SRC→M-TRG	X	50.0	46.7	51.0	73.9	72.7	73.5	61.9	59.7	62.2
M-TRG+SRC+HR	X	48.4	47.3	49.8	73.4	72.8	75.1	60.9	60.1	62.4
Translate-Test										
SRC	R	31.8	31.7	32.4	64.4	64.3	64.4	48.1	48.0	48.4
SRC	X	30.5	30.3	32.1	62.6	62.4	62.5	46.6	46.4	47.3
Roundtrip-Train-Test										
RT+SRC	R	30.4	29.7	31.7	61.9	62.1	62.9	46.2	45.9	47.3
RT+SRC	X	30.6	30.6	32.3	60.0	59.5	60.3	45.3	45.1	46.3
M-RT+SRC	R	30.8	30.8	34.1	59.3	58.7	59.4	45.0	44.7	46.8
M-RT+SRC	X	30.4	31.0	32.4	57.6	56.4	58.2	44.0	43.7	45.3
M-RT-Ens-SRC	X	34.5	35.4	35.4	57.9	57.3	58.9	46.2	46.3	47.2
M-RT-Ens-HR	X	35.4	35.1	37.1	50.2	49.3	50.3	42.8	42.2	43.7

Table 18: Results for translation-based XLT evaluated of Masakha for languages **not** supported by the translation model. Model selection is done on the best epoch based on source-language validation data (*Val-Src* (I)), based on translated source-language validation data (*Val-MT-Trg* (II)), and based on target-language validation data (*Val-Trg* (III)). We use XLM-R (X) and RoBERTa (R).