When Does Monolingual Data Help Multilingual Translation: The Role of Domain and Model Scale

Anonymous ACL submission

Abstract

Multilingual machine translation (MMT), trained on a mixture of parallel and monolingual data, is key for improving translation in low-resource language pairs. However, the literature offers conflicting results on the performance of different methods of including monolingual data. To resolve this, we examine how denoising autoencoding (DAE) and backtranslation (BT) impact MMT under different data conditions and model scales. Unlike prior studies, we use a realistic dataset of 100 translation directions and consider many domain combinations of monolingual and test data. We find that monolingual data generally helps MMT, but models are surprisingly brittle to domain mismatches, especially at smaller model scales. BT is beneficial when the parallel, monolingual, and test data sources are similar but can be detrimental otherwise, while DAE is less effective than previously reported. Next, we analyze the impact of scale (from 90M to 1.6B parameters) and find it is important for both methods, particularly DAE. As scale increases, DAE transitions from underperforming the parallel-only baseline at 90M to converging with BT performance at 1.6B, and even surpassing it in low-resource. These results offer new insights into how to best use monolingual data in MMT.

1 Introduction

The need for large supervised corpora remains a major bottleneck in neural machine translation (NMT) (Bapna et al., 2022). Sufficient bilingual data is scarce for most languages and limited to religious texts for the lowest-resource languages. To compensate for this lack of data, one effective approach is to leverage *related parallel data* from other languages via multilingual machine translation (MMT) that enables positive transfer from high-resource to low-resource languages (Aharoni et al., 2019; Arivazhagan et al., 2019). Additionally, we can use *monolingual data*, either through pretraining with denoising autoencoding(DAE; Conneau and Lample 2019; Liu et al. 2020a), or with backtranslation (BT; Sennrich et al., 2016). Driven by the success of these methods, recent works are converging toward a unified approach, that jointly trains MMT with monolingual data using auxiliary DAE objectives (Siddhant et al., 2022; Bapna et al., 2022; NLLB team et al., 2022) and/or BT.

However, the literature contains contradictory results about the effectiveness of these methods, particularly DAE. Early studies indicated combining MMT with DAE led to improvements across all settings (Wang et al., 2020; Siddhant et al., 2020). These studies, however, were limited in scope, as they only considered moderately-sized models and used few languages (10 to 15), with training and test data drawn from similar domains. By contrast, NLLB team et al. (2022) found that DAE helped only in very low-resource directions in MMT experiments with 200+ languages, while Xu et al. (2023) reported that DAE produced mixed results in experiments with (mostly) African languages.

To resolve this conflict, we present a systematic analysis of different methods that integrate monolingual data into MMT, focusing on BT and two DAE objectives, MASS (Song et al., 2019) and BART (Lewis et al., 2020; Liu et al., 2020b). First, we carefully investigate the role of the *domain*. We consider a realistic and diverse multilingual translation dataset with 100 directions, and run controlled experiments using different monolingual splits with single- and mixed-domain data. Then, we evaluate models across four wide-coverage multilingual test sets from Wikipedia, news, medical, and mixed domains. Our results with medium-sized models (370M) show that while BT outperforms both DAE objectives in most settings, the effectiveness of all methods varies significantly, as they are surprisingly brittle to domain mismatches. BT is more sensitive to the domain than DAE, and can underperform the parallel-only baseline when the monolingual and test data are not similar. However, increas-

ing the diversity of the monolingual data by mixing different sources improves domain robustness to some extent. We also discover that both DAE methods are less effective than previously reported, and they are mainly helpful in low-resource and xx→en directions. Of the two, MASS consistently outperforms BART, although by a narrow margin.

Next, we study the role of *model capacity* and discover that it is crucial and can even change the ranking between methods. We hold all other factors constant and train models with sizes from 90M up to 1.6B parameters. When the scale is small, both BT and DAE yield poor results, especially in out-of-domain settings. However, as model capacity grows, all methods quickly improve compared to the parallel-only baseline, and also become more robust to domain mismatches. Scale affects DAE the most, which transitions from underperforming the parallel-only baseline at the 90M scale to becoming competitive with BT at 1.6B and even outperforming it in low-resource.

Our contributions are: (i) We present a large-scale systematic analysis of how the *domain* and model *scale* affect the effectiveness of methods that incorporate monolingual data into MMT. (ii) We show that BT and DAE are sensitive to domain mismatches between the monolingual and test data, particularly on small scales. BT is best in most settings. Also, prior works have overestimated DAE, and when comparing the two methods, MASS outperforms BART. (iii) We discover that model capacity is key for the effectiveness of both methods, especially DAE. When the scale is small, DAE can even harm MMT, but it quickly improves with scale, and eventually becomes competitive with BT.

2 Related Work

Monolingual Data with Multi-Task Learning Early works on DAE+MMT report universal gains in all settings. Siddhant et al. (2020) use WMT parallel data from 15 languages and large monolingual corpora from many sources, like News Crawl, Wikipedia, and Common Crawl, with MASS. Wang et al. (2020) explore BART-like objectives with a subset of 10 languages from Siddhant et al. (2020) and News Crawl monolingual data.

However, more recent works that use larger and/or less uniform datasets, report less favourable results. To extend MMT to very low-resource languages, Bapna et al. (2022) show that models learn to translate from/into languages with only monolingual data if there are sufficient parallel data in other languages to enable transfer from the DAE to the MT task. NLLB team et al. (2022) explore a similar idea, but report that, in supervised translation, DAE (BART) is effective only for very low-resource. Xu et al. (2023) compare all aforementioned DAE methods and find that they often fail to outperform the parallel-only baseline. Our study probes confounding factors in these prior works.

Large Language Models Large language models (LLMs) trained on massive datasets achieve impressive results in many tasks (Brown et al., 2020; Chowdhery et al., 2022; Zhang et al., 2022b; Tay et al., 2023). To adapt LLMs to downstream tasks including translation (Wei et al., 2022; Lin et al., 2022; Zhang et al., 2023; Vilar et al., 2022; Garcia et al., 2023; Zhu et al., 2023; Hendy et al., 2023), the dominant approach is to use prompting, an ability enabled by model scale (Wei et al., 2022). Our work, however, is orthogonal and presents an analysis of methods that integrate monolingual data into encoder-decoder MMT models trained from scratch. Also, it is questionable whether these models are unsupervised with respect to translation, as recent work suggests that they have consumed parallel data during pretraining (Briakou et al., 2023).

Model Scale A growing literature investigates the scaling laws of different aspects of a model (Kaplan et al., 2020). In NMT, Ghorbani et al. (2021) explore scaling laws related to model capacity, Fernandes et al. (2023) consider MMT, and Gordon et al. (2021) focus on data scaling. Zhang et al. (2022a) investigate the scaling laws across architectures, like decoder-only and encoder-decoder. Our work does not study scaling laws but analyzes how scale impacts using monolingual data in MMT.

Analysis Huang et al. (2021); Liu et al. (2021) analyze the complementarity of BT and monolingual pretraining when used in bilingual NMT. By contrast, we focus on multilingual NMT and systematically analyze the joint training with BT and DAE.

3 (Multi-task) Multilingual NMT

We follow the universal MMT training method of Johnson et al. (2017) and train a single dense Transformer-based (Vaswani et al., 2017) model on the concatenation of parallel data from multiple language pairs. We prepend a special token ⟨2XX⟩ to the source sequences, that informs the model about the translation direction (e.g., ⟨2ES⟩ for Spanish).

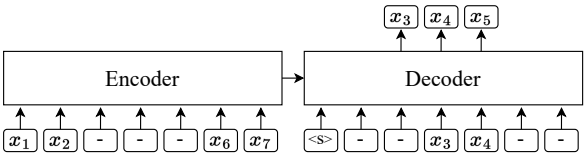


Figure 1: Illustration of the MASS objective.

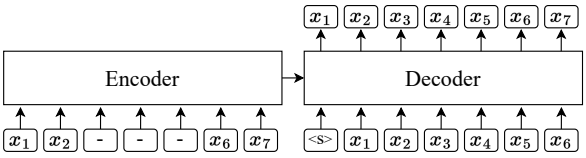


Figure 2: Illustration of the BART objective.

3.1 Denoising Autoencoding

We follow the multi-task setting from prior works (Siddhant et al., 2020; Wang et al., 2020) and use the regular MT objective on batches with parallel data and a DAE objective on batches with monolingual data. The language token ⟨2XX⟩ informs the model about the DAE and MT tasks, as it instructs it to generate a semantically similar sentence in the XX language. We explore two DAE methods.

MASS Song et al. (2019) adapt the masked language modeling objective (Devlin et al., 2019) to encoder-decoder models. MASS masks a span in the input and trains the decoder to predict that span. However, the unmasked tokens are not included in the target prefix (Figure 1). Following Siddhant et al. (2022, 2020), we do not use the architectural modifications of Song et al. (2019), such as extra language embeddings or custom initialization.

BART Lewis et al. (2020) propose a DAE objective similar to MASS, but with two differences. First, BART uses a slightly different noising strategy that can corrupt more than one input span in each sentence. Second, and more importantly, while the decoder is also trained to reconstruct the source sentence, its input context contains the full prefix, *including* the masked tokens (Figure 2).

3.2 Backtranslation

For BT, to save resources, instead of training separate bilingual models, we re-use the baseline MMT model and generate the new synthetic parallel data using the monolingual data of each language.

4 Experimental Setup

Parallel Data We use ML50 (Tang et al., 2021), a multilingual translation dataset between English and 50 other languages. ML50 is more representative of real-world multilingual datasets as it contains typologically diverse languages, including high, medium, and (extremely – less than 10k) low resource pairs, and with data from different domains. It is also more multilingual than the datasets from Siddhant et al. (2020) and Wang et al. (2020), that use 15 and 10 languages, respectively. To reduce training time, we cap the parallel data at 10M sentences per language, similar to Wang et al. 2020, which affects only few high-resource languages.

Monolingual Data We run *controlled* experiments with single- and mixed-domain monolingual data. For the single-domain experiments, we use Wikipedia as it is the only publicly available source with available data for all languages in ML50, but exclude the *xh* and *iu* languages from the experiments as they lack sufficient monolingual data. We cap the monolingual data per language to 5M, similar to Wang et al. (2020), which is still much larger than the parallel data for most languages. For the mixed-domain experiments, we use the *same* number of sentences per language, but also include News Crawl[1] (Barrault et al., 2020) and Web Crawl data from CC100[2] (Conneau et al., 2020). See the Appendix for the full data statistics (Table 16).

Evaluation Besides ML50 we also consider three *domain-specific* test sets. We use FLORES-200 (Goyal et al., 2022; NLLB team et al., 2022) with translations of *Wikipedia* articles, NTREX-128[3] (Federmann et al., 2022) with translations in 128 languages from the English WMT19 *News* test set (Barrault et al., 2019), and TICO-19 with translations in the *medical* domain (Anastasopoulos et al., 2020). FLORES-200 and NTREX-128 cover all languages in ML50, while TICO-19 covers only 15, but equally distributed across high, medium, and low resources. At test time, use beam search with K=5. In the main paper, we report results using BLEU (Papineni et al., 2002) similar to most prior works. However, to make our evaluation more comprehensive, we include in the Appendix the results from all experiments using ChrF (Popović, 2015) and COMET[4] (Rei et al., 2020), which is a neural metric. We find that overall all metrics are very consistent with each other, with few small differences in en→xx (see Appendix). We use SacreBLEU[5] (Post, 2018) for ChrF and BLEU.

[1] https://www.statmt.org/wmt20/translation-task.html
[2] https://data.statmt.org/cc-100/
[3] https://github.com/MicrosoftTranslator/NTREX. Because of misalignments, we omit the *ur* and *vi* languages.
[5] BLEU+case.mixed+lang.S-T+numrefs.1+smooth.exp+tok.13a+v1.5.1
[5] We use v2.0.1 with the *wmt22-comet-da* model.

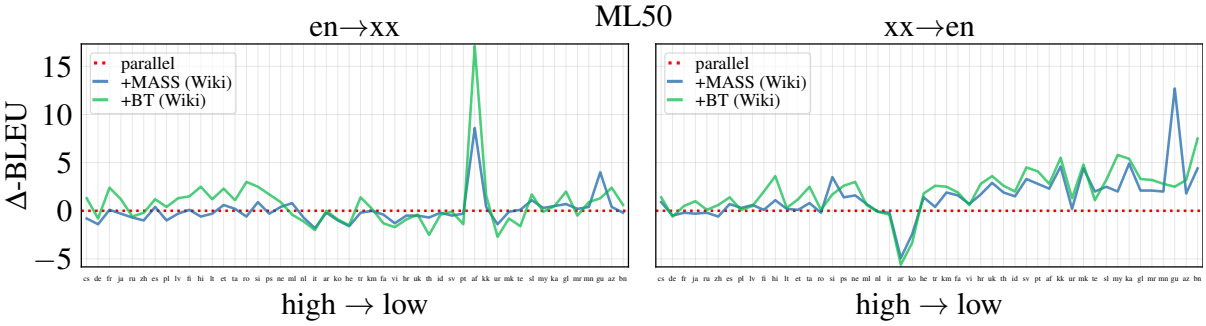


Figure 3: BLEU differences between each model and the parallel-only model (red dotted line) on the ML50 test data.

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
parallel	22.5	17.3	20.6	26.9	25.9	25.3	22.9
+BART	22.0	16.9	21.3	27.0	26.6	27.9	23.6
+MASS	22.1	16.9	21.3	27.1	26.5	28.5	23.7
+BT	**23.6**	**17.3**	**21.6**	**27.8**	**26.9**	**28.9**	**24.3**

Table 1: BLEU scores (↑) on the ML50 test. The models with BT and both SSL objectives (BART, MASS) use the single-domain monolingual split with data only from Wikipedia. The cells in red indicate that a model fails to improve over the parallel-only baseline.

Data Sampling We use temperature-based data sampling (Arivazhagan et al., 2019) to balance the training data. Assuming that p_{D} is the probability that a sentence belongs to dataset D, we sample sentences for D with a probability proportional to $p_{\mathrm{D}}^{1/T}$, where T is a temperature parameter. When using parallel data, D corresponds to the data of a given language pair. When including monolingual (i.e., for DAE) or synthetic parallel (i.e., for BT) data, we *first* concatenate all the separate datasets to the same list and *then* apply temperature sampling. That is, the real en→fr, synthetic (BT) en′ →fr, and monolingual fr↔fr, are treated as separate datasets D. Larger values of T lead to more even sampling (i.e., upsampling small datasets). We set $T = 5$ following prior works (Wang et al., 2020; Siddhant et al., 2020), which also leads to a roughly 1:1 ratio when using both monolingual and parallel data.

Models Our baseline is an MMT model trained *only* on the en→xx and xx→en parallel data. For both MASS and BART, we mask 50% of input tokens following the hyperparameters from Siddhant et al. (2022, 2020) and NLLB team et al. (2022), respectively. All models use the same Transformer architecture (Vaswani et al., 2017). We consider three different model sizes for our scaling experiments: 1) *Transformer-Base* with 90M parameters, 2) *Transformer-Big* with 370M parameters, and 3) *Transformer-XL* (not to be confused with Dai et al. 2019), with 1.6B parameters. We include details about our models and training in Appendix A.

5 Results

5.1 Single-Domain Monolingual Data (Wiki)

We begin with a series of controlled experiments that measure the impact of the domain using the Transformer-Big model scale (370M). We compare across different test sets the parallel-only model with parallel+BT and parallel+DAE (MASS, BART) that use the single-domain monolingual split (see statistics in Table 16). In Table 1 we report the BLEU scores of each model on the ML50 test set averaged by group and translation direction.

On average, BT and both DAE models outperform the baseline by +1.4 and +0.7 BLEU points, respectively. BT consistently achieves the best results, with the largest gains in low-resource, with +1 BLEU points on en→xx and +3.6 BLEU points on xx→en. Both DAE models produce similar results, but MASS is marginally better. However, in the en→xx high- and medium-resource languages, both DAE models fail to outperform the baseline, although they use the same monolingual data as BT.

Non-aggregated scores reveal mixed results. To get a more detailed picture of model performance we plot the *differences* in the BLEU scores (Δ-BLEU) between each model and the parallel-only baseline model across all pairs in Figure 3. For a simpler presentation, we omit BART which is similar to MASS. Figure 3 reveals that the results are more mixed than the aggregated scores suggest (Table 1). In xx→en, both BT and MASS are generally better than the baseline and follow a similar trend. Their gains increase towards the low-resource languages, with few exceptions, and BT is better than MASS in most cases. However, in en→xx, we discover a different picture. BT shows

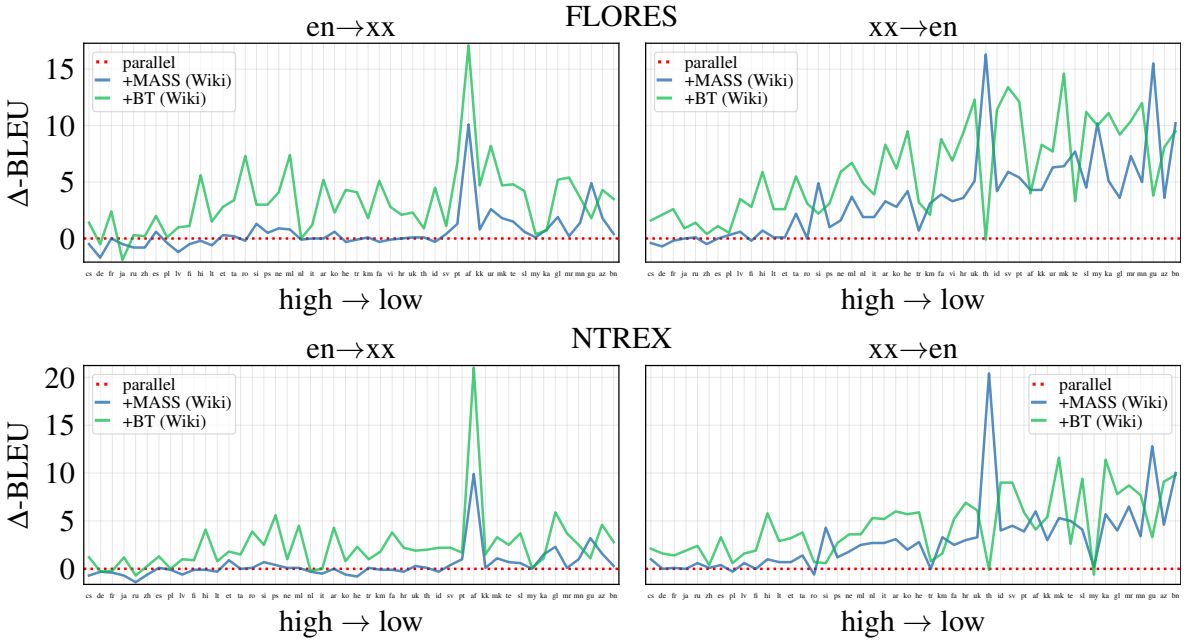


Figure 4: BLEU differences between each model and the baseline (red dotted line) on FLORES and NTREX.

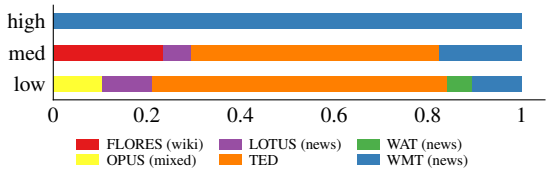


Figure 5: Data sources used for the ML50 test sets.

a surprising behavior as it outperforms the baseline in high-resource (usually from +2 to +4 BLEU) but harms BLEU in most medium- to low-resource languages and is also often worse than MASS. MASS fluctuates around the baseline and benefits only a few low-resource languages. These results contradict early works on MMT+DAE that report *universal* gains (Siddhant et al., 2020; Wang et al., 2020).

What is the reason for the mixed results? In our experiments, we used the same model/training hyperparameters as in previous conflicting studies (Wang et al., 2020; Siddhant et al., 2020). The only difference lies in our training and test data. Those earlier works used 10/15 languages from WMT and news test sets. By contrast, the ML50 dataset is more challenging, as 1) it has more languages, 2) contains truly low-resource languages (24/50 have less than 200K sentences, unlike prior works), and, more importantly, 3) it has data from diverse sources (Figure 5). High-resource languages contain WMT (news) data, whereas other languages have data from different sources, mainly from TED talks. Recall that BT is more effective in high-resource pairs but yields poor results in non-English non-WMT pairs. Considering this, we hypothesize that previous works reported universal gains because they considered more favourabe experimental setups, with fewer languages and parallel, monolingual, and test data in the same domain.

How do results change on other test domains? To test this hypothesis, we evaluate models on *uniform* test sets, where all languages have data from the same source. Figure 4 shows the results on the FLORES (Wikipedia domain) and NTREX (news domain) test sets. The TICO-19 results follow similar trends and include them in the appendix.

The results in both FLORES and NTREX reveal a more favorable picture for both methods. We see similar trends as in the ML50 test sets, especially in xx→en, but the gains are overall larger. This can be explained by the greater domain similarity of the test sets with the monolingual data, particularly FLORES, which shows the biggest improvements. The switch to the in-domain test sets has a stronger effect on BT, especially in en→xx. Notice that in ML50, BT is harmful in en→xx low-resource with mostly out-of-domain data, whereas in NTREX and FLORES, it is consistently helpful. MASS also performs much better on the in-domain test data. However, we still fail to observe the universal gains reported in some works. For instance, in en→xx, it outperforms the baseline only in low-resource. We hypothesize that DAE requires more ideal conditions to be helpful in MMT. For instance, Siddhant et al. (2020) used much more monolingual

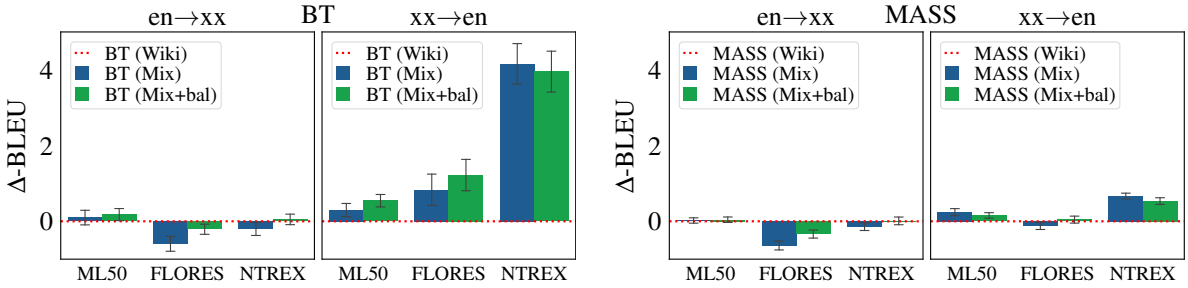


Figure 6: BLEU differences (Δ-BLEU) of the BT models trained with the mixed-domain split with respect to the single-domain monolingual data (dotted red line). To plot the bars, we use the mean Δ-BLEU and the standard error.

relative to the parallel data, whereas Wang et al. (2020) used a similar ratio to this work but with parallel, monolingual and test data from the same domain. Overall, the performance gap between test sets shows that the domain of the monolingual data is crucial and that both methods are sensitive to mismatches with the test domain, particularly BT.

5.2 Mixed-domain Monolingual Data

Previously, we examined single-domain monolingual data, removing confounding factors to isolate domain impact. We now turn to a real-world scenario and use multiple sources of monolingual data per language. The goal is to evaluate the significance of diversity in monolingual data. For each language, we hold the size of monolingual data constant (§5.1), and *only* change the data mixture. We include data from News Crawl and CC100 (web domain), the only other publicly available data sources with wide enough coverage to support most languages in ML50. For languages that do not have data from all domains, we use only the available ones. We consider two mixed-domain splits:

1. *Unbalanced:* This split emulates naively concatenating all the monolingual data of a given language without considering their relative sizes. The ratio between sources is proportional to the size of their uncapped data.
2. *Balanced:* This split balances the number of sentences from each source using the same temperature-based sampling method applied to the parallel data, with T=5.

In Figure 6, each bar shows the average BLEU difference (Δ-BLEU) compared to the single-domain split (Wiki). We include results on the TICO-19 and with ChrF scores in the appendix. Diversity largely favours BT with a minor impact on MASS. This further supports that BT is more sensitive to the domain. BT displays a contrast between translation directions. Note that 1) both BT and DAE use identical target-side monolingual data, and 2) the MMT model has been exposed to a large number of diverse (i.e., many domains) English target-side sentences through the ML50 parallel data. Thus, we hypothesize that source-side diversity causes the xx→en gains of BT.

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
FLORES							
BART	23.6	15.2	14.3	27.8	24.4	22.0	20.8
MASS	**23.8**	**15.3**	**14.5**	**28.4**	**25.0**	**23.4**	**21.3**
NTREX							
BART	21.8	13.1	**13.7**	25.0	23.2	21.0	19.4
MASS	**21.9**	**13.3**	**13.7**	**25.7**	**23.8**	**22.1**	**19.8**
ML50							
BART	**22.1**	**16.8**	21.3	26.8	26.1	28.1	23.5
MASS	**22.1**	**16.8**	**21.5**	**27.2**	**26.6**	**28.8**	**23.8**
TICO-19							
BART	31.2	14.0	15.1	31.8	26.0	24.1	23.7
MASS	**31.5**	**14.4**	**15.2**	**32.6**	**27.2**	**26.4**	**24.5**

Table 2: BLEU scores (↑) of BART and MASS trained with the balanced mixed-domain monolingual data.

The highest gains appear in NTREX test sets (up to +4 BLEU), as mixed splits incorporate monolingual data from the same domain, i.e., news. Interestingly, mixed-domain data proves beneficial for xx→en in FLORES. Closer examination reveals that these gains mainly affect low-resource languages (Table 5). Although the reason isn't clear, we speculate it may be due to reduced cross-domain interference between the parallel and monolingual data. The re-balancing of monolingual data has minimal impact, though it does slightly enhance or mitigate the drawbacks of using less in-domain data (e.g., FLORES). NTREX does not benefit because re-balancing leads to using less news data.

5.3 Denoising Autoencoding Objectives

Table 2 compares MASS and BART across all test sets. We consider their variants trained with the balanced monolingual data (§5.2), as they work marginally better (see Appendix §B.2 for more re-

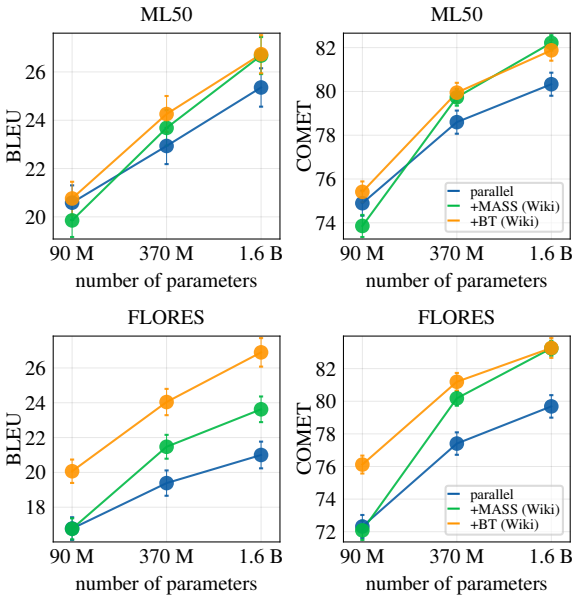


Figure 7: Mean BLEU and COMET across model scales. The error bars show the standard error of the mean.

sults). MASS consistently outperforms BART, with larger gains in xx→en (up to 2 BLEU). However, in xx→en, their results are comparable.

Both objectives use similar encoder noising methods but differ in the decoder. BART's decoder conditions on the full target prefix, unlike MASS, which excludes unmasked tokens. This potentially makes the MASS decoder rely more on its encoder. Next, BART computes loss over all tokens, even unmasked ones, consequently losing part of the useful signal by teaching the model to copy the input. MASS, however, calculates loss only on unmasked tokens, targeting the training signal to denoising. In related work, Baziotis et al. (2021) study NMT pretraining using BART variants with different input noising methods, such as word replacement or shuffling, and present evidence that input masking biases models towards copying the input. We conclude that the performance gap between MASS and BART stems from these decoder-side differences.

5.4 Scale

This section examines the role of model scale. We hold all other factors constant and test three model sizes that differ by a factor of 4: Transformer-Base (90M), Transformer-Big (370M), and Transformer-XL (1.6B). To conserve computational resources, we consider only one DAE method, MASS, as it outperformed BART in previous experiments. We use the (Wiki) single-domain monolingual split to test for in-domain (FLORES) and out-of-domain (ML50) effects. Figure 7 shows results and includes BLEU and COMET[6].

How crucial is model capacity for BT and DAE? All models improve with scale. However, small models find monolingual methods less beneficial, especially in ML50 (top), which is out-of-domain with respect to the (Wikipedia) monolingual data. BT shows negligible gains, while MASS even proves detrimental. As scale increases, both MASS and BT become more effective, with MASS benefiting the most. Surprisingly, MASS transitions from underperforming the baseline to outperforming it and becomes competitive with BT at the 1.6B scale. We also discover that according to COMET (and chrF), the effects of scale on MASS are even stronger, as it outperforms BT by a small margin.

In FLORES (bottom), BT and MASS exhibit a similar trend, but are overall more effective, since the test and monolingual domains are the same. At small scale, MASS fails to yield any gains, whereas BT is more helpful. As scale increases, the gains of both methods relative to the baseline also increase. However, according to BLEU, the performance gap between MASS and BT remains relatively constant, unlike in ML50, whereas according to COMET, MASS achieves again comparable performance to BT. This suggests that DAE becomes more competitive with scale and bridges the gap with BT, in particular in out-of-domain settings (ML50).

We speculate that learning from monolingual data proves more challenging for smaller models because they prioritize learning from parallel data. This also explains why BT outperforms DAE at small scales. Translating the synthetic parallel data, which is more similar to the supervised MT task, is an easier task compared to denoising. As model capacity increases, it "unblocks" DAE and progressively enables it to make better use of monolingual data. This suggests that there is a cross-task interference that is mitigated by scaling.

How direction and resource-level are affected? Next, we investigate the scaling patterns of MASS and BT. Figure 8 shows the relative difference between the BLEU score of each model and the corresponding parallel-only baseline in the same scale across translation directions. Both methods benefit from scale, with low-resource settings gaining the most. Notice that for each method, that gap

[6]We include COMET here because, whilst in other experiments COMET and BLEU show similar results, in this case, we discover a small but noteworthy difference (see Appendix for details; Figures 12, 13).

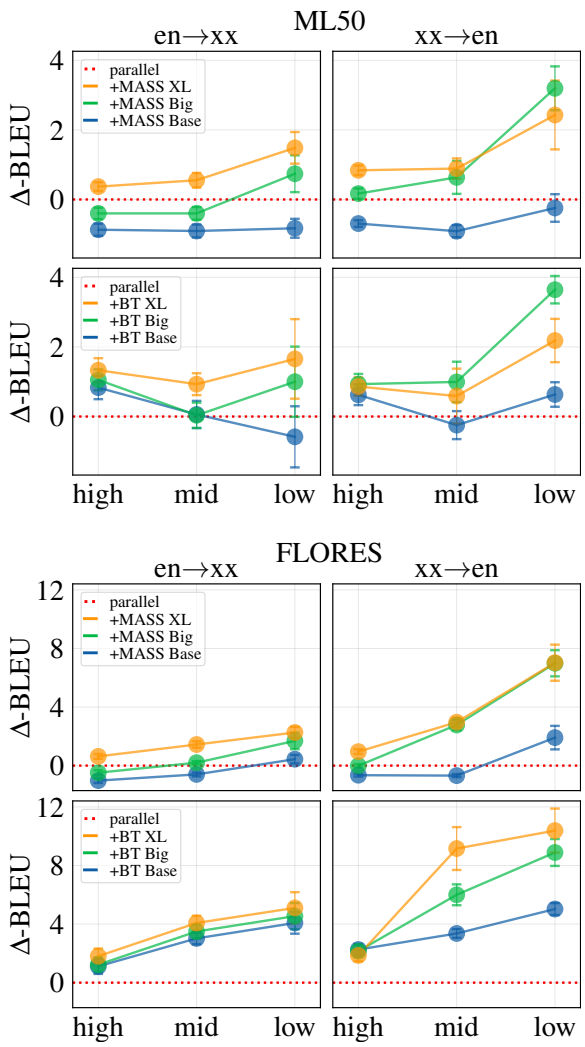


Figure 8: Average BLEU differences (Δ-BLEU) of each model with respect to the corresponding parallel-only baseline in the same scale (red dotted line). The error bars show the standard error of the mean.

between scales is small in high-resource (up to 2 BLEU) but large in low-resource directions (up to 3 and 5 BLEU in ML50 and FLORES, respectively). Scale also generally benefits more xx→en (right side) compared to en→xx (left side). The plots per test set also have the same y-axis, which enables us to directly compare BT with MASS. We discover that the reason MASS (on average) closes the gap with BT (see Figure 7) as scale grows is because of its low-resource performance. In particular, in ML50 at the 1.6B scale, the gap becomes negligible, and MASS even marginally outperforms BT in low-resource xx→en (two top-right plots).

6 Conclusion

This work presents a systematic analysis of widely used methods that include monolingual data in MMT, specifically BT, and two DAE objectives. It does not negate findings from prior works but rather highlights confounding factors that explain the mixed results found in the literature. These factors range from the characteristics of the experimental setup, like the data mixture, to the effective model capacity. The main takeaway is that one should not expect gains from DAE or BT in all settings but carefully consider all aspects of the system to reach optimal performance.

We compare models across different data conditions and combinations of monolingual and test data, and discover that all methods are very sensitive to domain mismatches. BT overall yields the most gains, but it can fail in out-of-domain and low-resource settings. As for DAE, we conclude that it can be helpful, particularly in low-resource and xx→en, but the universal gains reported from early works can only be achieved in ideal conditions, where the parallel, monolingual, and test data are from the same domain. Another key finding is that model capacity can make or break a method. Larger models are better able to use monolingual data, with gains from both BT and DAE increasing as the model scale grows. We also discover a novel connection between domain robustness and model size. Scale is more important in out-of-domain settings, as all methods yield limited to no gains at small scales. In particular, MASS is harmful to MMT with the 90M models, but when using 1.6B models, it becomes comparable or even better to BT.

Based on our findings, we provide some recommendations to practitioners:

- For in-domain settings, prefer BT, as it yields the best results across scales and resource levels.
- For out-of-domain settings, the choice depends on model size. At small scales, prefer BT but expect small gains. At large scales, both methods are more effective, and the gap between them diminishes. DAE is a viable and computationally cheaper alternative to BT, which needs to back-translate monolingual data from many languages.
- For MMT+DAE, prefer MASS instead of BART.
- Aim to increase the diversity of the monolingual data by mixing different sources and re-balance them to ensure a more even distribution.
- If in-domain or diverse monolingual data is not available, consider the trade-offs between collecting extra data or scaling up the model. If neither is possible, avoid using monolingual data with BT or DAE in en→xx low-resource directions.

Limitations

We used only one dataset with roughly 200M sentences and 100 translation directions. The dataset is more diverse, with more languages than many prior works, however, it is unclear how the results will generalize to datasets with other characteristics, such as more languages or more/less typologically diverse languages. The same holds for the combinations of monolingual and test data. We consider three main sources of publicly available monolingual data that also have wide coverage across many languages. Using more domains for the monolingual and test data would be better, but we could not find other monolingual sources with wide coverage.

Another limitation is that we only consider three model sizes (90M, 370M, and 1.6B). Different results might be obtained from smaller or larger models. We were particularly interested in whether DAE would outperform BT if we scaled models to over 1.6B parameters. However, this would require significantly more resources than we had available.

Lastly, we focused on the three most widely adopted methods for integrating monolingual data into MMT, namely BT and DAE with MASS/BART. However, there are other methods, such as those using contrastive losses (Pan et al., 2021). We leave these comparisons for future work.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. 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 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Christos Baziotis, Ivan Titov, Alexandra Birch, and Barry Haddow. 2021. Exploring unsupervised pretraining objectives for machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2956–2971, Online. Association for Computational Linguistics.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in palm's translation capability. *arXiv preprint arXiv:2305.10266*.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

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 and Guillaume Lample. 2019. Cross-lingual language model pretraining. In H Wallach, H Larochelle, A Beygelzimer, F d\Alché-Buc, E Fox, and R Garnett, editors, *Advances in Neural Information Processing Systems*, pages 7057–7067. Curran Associates, Inc.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. 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.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Patrick Fernandes, Behrooz Ghorbani, Xavier Garcia, Markus Freitag, and Orhan Firat. 2023. Scaling laws for multilingual neural machine translation. *arXiv preprint arXiv:2302.09650*.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. *arXiv preprint arXiv:2302.01398*.

Behrooz Ghorbani, Orhan Firat, Markus Freitag, Ankur Bapna, Maxim Krikun, Xavier Garcia, Ciprian Chelba, and Colin Cherry. 2021. Scaling laws for neural machine translation. *arXiv preprint arXiv:2109.07740*.

Mitchell A Gordon, Kevin Duh, and Jared Kaplan. 2021. Data and parameter scaling laws for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5915–5922, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Dandan Huang, Kun Wang, and Yue Zhang. 2021. A comparison between pre-training and large-scale back-translation for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1718–1732, Online. Association for Computational Linguistics.

Hakan Inan, Khashayar Khosravi, and Richard Socher. 2017. Tying word vectors and word classifiers: A loss framework for language modeling. In *Proceedings of the International Conference on Learning Representations*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, San Diego, CA, USA.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. 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.

Xuebo Liu, Longyue Wang, Derek F. Wong, Liang Ding, Lidia S. Chao, Shuming Shi, and Zhaopeng Tu. 2021. On the complementarity between pre-training and back-translation for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2900–2907, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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

NLLB team, Marta Costa-jussa, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janicec Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Gonzalez, Prangthip Hansanti, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. 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 244–258, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2827–2835, Online. Association for Computational Linguistics.

Aditya Siddhant, Ankur Bapna, Orhan Firat, Yuan Cao, Mia Xu Chen, Isaac Caswell, and Xavier Garcia. 2022. Towards the next 1000 languages in multilingual machine translation: Exploring the synergy between supervised and self-supervised learning. *arXiv preprint arXiv:2201.03110*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936, Long Beach, California, USA. PMLR.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. UL2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 5998–6008, Long Beach, CA, USA.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2022. Prompting palm for translation: Assessing strategies and performance. *arXiv preprint arXiv:2211.09102*.

Yiren Wang, ChengXiang Zhai, and Hany Hassan. 2020. Multi-task learning for multilingual neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1022–1034, Online. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Haoran Xu, Jean Maillard, and Vedanuj Goswami. 2023. Language-aware multilingual machine translation with self-supervised learning.

Biao Zhang, Behrooz Ghorbani, Ankur Bapna, Yong Cheng, Xavier Garcia, Jonathan Shen, and Orhan Firat. 2022a. Examining scaling and transfer of language model architectures for machine translation. In *International Conference on Machine Learning*, pages 26176–26192. PMLR.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. *arXiv preprint arXiv:2301.07069*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022b. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

A Experimental Setup

A.1 Training

Our baseline is an MMT model trained only on the parallel data (en→xx and xx→en). For BT, we use the baseline model to generate the synthetic translations using beam search with beam size 4, following NLLB team et al. (2022). For MASS, we use the hyperparameters from Siddhant et al. (2022, 2020) and mask 50% of input tokens. For BART, we use the hyperparameters[7] from NLLB team et al. (2022), that also mask 50% of input tokens. We implement all our models using the fairseq toolkit (Ott et al., 2019), and for BART we use the original implementation in fairseq, whereas for MASS develop our own re-implementation.

All models use the same Transformer architecture (Vaswani et al., 2017) with shared encoder-decoder embeddings and decoder output projection layers (Press and Wolf, 2017; Inan et al., 2017) as in NLLB team et al. (2022). We optimize our models with Adam (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-6}$, with a learning rate of 0.001 using a linear warm-up of 8k steps, followed by inverted squared decay. We also regularize the models with label smoothing (Szegedy et al., 2016) of 0.1 and weight decay of 0.01.

We consider three different model sizes: 1) *Transformer-Base* with 90M parameters configured as in the original paper, 2) *Transformer-Big* with 370M parameters, similar to the original but with an 8192-sized feed-forward layer as in Wang et al. (2020); Siddhant et al. (2020), and 3) *Transformer-XL* (not to be confused with Dai et al. (2019)), with 1.6B parameters, 12 encoder/decoder layers, feed-forward layers of 8192, 2048-sized embeddings, and 32 attention heads. We train all models with mixed precision (FP16) and use gradient accumulation to reach the desired batch size for each model size. Specifically, we train the Transformer-Base on 4 A100 GPUs for 440K steps with an effective batch size of 280K token batches, the Transformer-Big on 8 A100 GPUs for 360K steps with 320K token batches, and the Transformer-XL on 12 A100 GPUs for 120K steps with 860K token batches. We evaluate models every 40K (10k for Transformer-XL) steps and select the checkpoint with the best average translation loss (i.e., negative log-likelihood) across all language pairs in the ML50 validation set.

[7]Fairseq arguments: "–mask 0.5 –mask-random 0.1 –mask-length span-poisson –poisson-lambda 3.5"

	Base	Big	XL
Parameters (Size)	90M	370M	1.6B
Layers	6	6	12
Embedding	512	1024	2048
FeedForward	2048	8192	8192
Heads	8	16	32
Effective batch size	280K	320K	860K
Training Steps	440K	360K	120K
Dropout	0.1	0.3	0.3
GPU Configuration	4×A100	8×A100	12×A100

Table 3: Hyperparameters used for the Transformer models of various sizes in the study.

B Additional Results

In the main paper, for brevity, we discuss results using only BLEU and for selected experiments that highlight our most important findings. For completeness, we also re-evaluate the outputs from *all* of our experiments and across all test sets with two additional evaluation metrics, following the recommendations of Kocmi et al. (2021):

chrF: this is another surface-level (i.e., string-based) metric, like BLEU, but achieves better correlation with human judgment. It compares character n-grams that make it better for languages with rich morphology and is also tokenization independent.

COMET: this is a neural-based metric that uses a pretrained model to estimate the translation quality. Unlike BLEU and chrF, it also takes into account the source sentence. However, we point out that it is not clear how reliable (the current version of) COMET is for low-resource languages or test data across different domains, as Kocmi et al. (2021) in their analysis considered only high-resource languages and two test domains (news, discussions).

We find that overall, the *ranking* of the models is very consistent across metrics. We observe only two instances where metrics do not fully agree with each other, mainly in en→xx and low-resource languages (see §B.1.1, §B.3). However, the main findings and patterns discussed in the main body of the paper still hold across metrics.

B.1 Main Experiments

First, we report the results of the experiments that investigate the role of data. This includes the results from all models trained with the single-domain (Wikipedia) and mixed-domain (unbalance-vs-balanced) monolingual data in Section 5.1 and Section 5.2, respectively. Recall that ML50 contains parallel data from many different sources,

which are mostly out-of-domain data with respect to the Wikipedia domain. The same holds for the ML50 test data.

We include the full results for all methods across all monolingual splits in Table 4 (ML50), Table 5 (FLORES), Table 6 (NTREX) and Table 7 (TICO-19). Next, we also include the line charts with the score differences of all models with all metrics in Figure 8, which are the counterparts of the Figures 3, 4 in the main body of the paper.

B.1.1 Mixed-Domain Monolingual Data

Besides the table view of the results, which do include the scores per monolingual split, here we also report the corresponding bar plots, similar to those in Section 5.2, with all methods, test sets, and metrics. This is one of the few cases where we discover a small discrepancy between metrics. Specifically, we see that the ChrF and COMET results suggest that using mixed-domain monolingual data is even *more helpful* for BT, than what the BLEU scores suggest. In particular, Figure 9 shows gains in BLEU (top) only in the xx→en direction, whereas the ChrF (middle row) and COMET (bottom row) scores reveal consistent improvements even in the en→xx direction. We also see that further re-balancing (green bar) the monolingual data yields small gains in most settings. Besides these differences, the overall trends are the same across metrics (i.e., BT is more sensitive to diversity than MASS, with larger gains in xx→en).

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
parallel	22.5	17.3	20.6	26.9	25.9	25.3	22.9
Wiki							
+BART	22.0	16.9	21.3	27.0	26.6	27.9	23.6
+MASS	22.1	16.9	21.3	27.1	26.5	28.5	23.7
+BT	23.6	17.3	**21.6**	27.8	26.9	28.9	24.3
Mix							
+BART	21.8	16.7	21.4	27.1	26.3	28.4	23.6
+MASS	22.0	16.8	21.5	27.4	26.5	28.9	23.8
+BT	24.0	**17.5**	21.3	28.3	26.9	29.4	24.5
Mix+bal							
+BART	22.1	16.8	21.3	26.8	26.1	28.1	23.5
+MASS	22.1	16.8	21.5	27.2	26.6	28.8	23.8
+BT	**24.1**	**17.5**	21.4	**28.5**	**27.2**	**29.6**	**24.6**

(a) BLEU scores (↑)

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
parallel	48.6	45.0	46.3	55.3	49.3	47.0	48.3
Wiki							
+BART	47.9	44.6	47.4	55.1	50.5	50.4	49.1
+MASS	48.1	44.5	47.5	55.2	50.6	50.9	49.3
+BT	49.7	45.4	47.2	56.4	51.6	51.7	50.1
Mix							
+BART	47.7	44.0	47.4	55.3	50.4	50.8	49.1
+MASS	47.9	44.4	47.5	55.3	50.4	51.2	49.3
+BT	49.8	**46.1**	**48.0**	56.3	51.6	52.4	**50.5**
Mix+bal							
+BART	47.9	44.1	47.4	55.1	50.4	50.6	49.1
+MASS	48.0	44.4	47.6	55.2	50.6	51.2	49.3
+BT	**50.0**	45.7	47.8	**56.5**	**51.7**	**52.5**	**50.5**

(b) chrF scores (↑)

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
parallel	80.9	80.5	77.0	80.8	78.0	75.6	78.6
Wiki							
+BART	80.4	80.2	78.4	80.9	79.1	78.9	79.5
+MASS	80.7	79.9	78.6	81.0	79.4	79.4	79.7
+BT	81.8	80.9	78.3	81.3	79.4	78.9	80.0
Mix							
+BART	80.0	79.2	78.7	80.9	79.0	79.3	79.4
+MASS	80.5	79.9	**78.8**	81.2	79.3	**79.8**	79.8
+BT	82.2	**81.6**	78.7	81.4	79.4	79.6	80.3
Mix+bal							
+BART	80.2	79.5	78.3	80.8	78.9	79.0	79.4
+MASS	80.6	79.7	**78.8**	81.2	79.3	**79.8**	79.8
+BT	**82.3**	81.2	**78.8**	**81.6**	**79.6**	**79.8**	**80.4**

(c) COMET scores (↑)

Table 4: Results of the Transformer-Big models evaluated on the **ML50** (mixed-domain) test set and grouped by the monolingual split that has been used for training BT and DAE.

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
parallel	24.8	15.3	13.2	28.6	22.4	16.0	19.4
Wiki							
+BART	24.0	15.6	14.7	28.3	24.9	22.5	21.2
+MASS	24.3	15.5	14.9	28.6	25.2	23.0	21.5
+BT	**26.0**	**18.8**	17.7	30.8	28.4	24.9	24.1
Mix							
+BART	23.4	15.0	14.2	27.9	24.7	22.6	20.9
+MASS	23.5	15.2	14.1	28.5	24.8	23.0	21.1
+BT	25.6	17.6	17.6	30.8	28.6	26.9	24.2
Mix+bal							
+BART	23.6	15.2	14.3	27.8	24.4	22.0	20.8
+MASS	23.8	15.3	14.5	28.4	25.0	23.4	21.3
+BT	25.5	18.3	**18.0**	**31.1**	**29.1**	**27.2**	**24.6**

(a) BLEU scores (↑)

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
parallel	50.6	46.1	45.0	57.1	50.4	42.6	48.2
Wiki							
+BART	49.9	46.2	47.1	56.9	53.1	50.6	50.4
+MASS	50.2	46.1	47.3	57.0	53.4	51.3	50.6
+BT	**52.1**	**49.3**	47.6	59.2	57.1	52.2	52.7
Mix							
+BART	49.5	45.1	46.4	56.5	52.9	50.6	49.9
+MASS	49.4	45.7	46.4	56.7	53.0	51.2	50.1
+BT	51.6	49.2	**49.3**	59.1	57.0	55.1	53.4
Mix+bal							
+BART	49.4	45.3	46.5	56.4	52.9	50.4	49.9
+MASS	49.7	45.8	46.8	56.8	53.3	51.6	50.4
+BT	51.7	49.2	**49.3**	**59.3**	**57.3**	**55.2**	**53.5**

(b) chrF scores (↑)

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
parallel	83.0	80.5	71.4	84.0	79.3	69.9	77.4
Wiki							
+BART	82.5	80.5	74.4	83.9	81.5	78.9	80.0
+MASS	82.8	80.3	74.6	83.9	81.7	79.6	80.2
+BT	84.1	82.7	77.5	**84.8**	82.9	77.2	81.2
Mix							
+BART	81.8	79.1	73.5	83.4	81.1	78.6	79.3
+MASS	82.0	80.0	73.7	83.7	81.3	79.3	79.7
+BT	84.1	**83.0**	78.1	84.6	82.7	80.1	81.9
Mix+bal							
+BART	81.9	79.5	73.4	83.3	81.0	78.5	79.3
+MASS	82.3	80.0	74.0	83.7	81.6	79.7	79.9
+BT	**84.2**	82.9	**78.6**	**84.8**	**83.0**	**80.4**	**82.1**

(c) COMET scores (↑)

Table 5: Results of the Transformer-Big models on the **FLORES** (Wikipedia) test set and grouped by the monolingual split that has been used for training BT and DAE. Cells in red indicate worse scores than the baseline.

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
parallel	22.4	13.2	12.4	25.1	21.1	15.1	17.8
Wiki							
+BART	21.9	13.2	13.4	25.5	23.3	20.8	19.4
+MASS	22.1	13.2	13.8	25.5	23.3	21.2	19.5
+BT	**23.3**	**15.5**	16.0	27.4	25.1	21.8	21.2
Mix							
+BART	21.6	13.0	13.6	25.4	23.4	21.5	19.5
+MASS	21.7	13.1	13.7	26.0	23.9	22.1	19.8
+BT	22.8	14.9	16.4	**30.9**	**28.6**	**27.1**	**23.2**
Mix+bal							
+BART	21.8	13.1	13.7	25.0	23.2	21.0	19.4
+MASS	21.9	13.3	13.7	25.7	23.8	22.1	19.8
+BT	22.9	15.4	**16.6**	30.4	28.5	27.0	**23.2**

(a) BLEU scores (↑)

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
parallel	48.3	43.3	42.2	54.5	49.3	40.8	46.0
Wiki							
+BART	47.6	43.2	44.0	54.5	51.7	47.9	47.9
+MASS	47.9	43.0	44.2	54.6	51.9	48.5	48.1
+BT	**49.2**	45.7	44.2	56.7	54.6	48.3	49.5
Mix							
+BART	47.4	42.8	43.9	54.5	51.7	48.4	47.9
+MASS	47.5	43.2	44.1	54.7	52.0	49.1	48.2
+BT	48.8	46.2	**46.2**	**58.7**	**56.6**	**53.1**	**51.4**
Mix+bal							
+BART	47.4	42.9	44.1	54.4	51.8	48.2	47.9
+MASS	47.6	43.3	44.3	54.6	52.1	49.3	48.3
+BT	49.1	**46.3**	**46.2**	58.4	56.5	52.8	**51.4**

(b) chrF scores (↑)

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
parallel	79.0	76.9	69.4	82.1	78.7	67.6	75.2
Wiki							
+BART	78.3	76.7	72.0	82.1	80.6	75.5	77.3
+MASS	78.8	76.4	72.3	82.3	80.8	76.3	77.6
+BT	79.7	78.7	74.4	83.0	81.6	73.4	78.2
Mix							
+BART	78.0	76.0	72.0	81.9	80.5	76.2	77.2
+MASS	78.5	76.9	72.3	82.3	81.0	76.9	77.8
+BT	80.1	79.6	76.2	**83.8**	82.6	**77.6**	79.8
Mix+bal							
+BART	78.2	76.3	71.9	81.8	80.5	75.9	77.2
+MASS	78.6	76.8	72.3	82.2	81.0	77.2	77.8
+BT	**80.2**	**79.7**	**76.3**	**83.8**	**82.7**	**77.6**	**79.9**

(c) COMET scores (↑)

Table 6: Results of the Transformer-Big models on the **NTREX** (News) test set and grouped by the monolingual split that has been used for training BT and DAE. Cells in red indicate worse scores than the baseline.

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
parallel	32.3	14.3	14.4	32.4	24.2	17.4	22.3
Wiki							
+BART	31.9	14.9	15.1	32.9	26.3	24.2	24.2
+MASS	31.9	14.0	15.4	32.9	27.0	24.6	24.3
+BT	**34.5**	**18.4**	19.8	36.8	32.2	28.7	28.3
Mix							
+BART	30.5	13.9	14.9	32.5	26.6	24.2	23.7
+MASS	31.1	14.3	15.2	33.0	26.9	25.6	24.3
+BT	33.2	16.5	19.2	36.9	32.5	30.6	28.2
Mix+bal							
+BART	31.2	14.0	15.1	31.8	26.0	24.1	23.7
+MASS	31.5	14.4	15.2	32.6	27.2	26.4	24.5
+BT	34.3	17.7	**20.2**	**37.4**	**33.0**	**30.9**	**28.9**

(a) BLEU scores (↑)

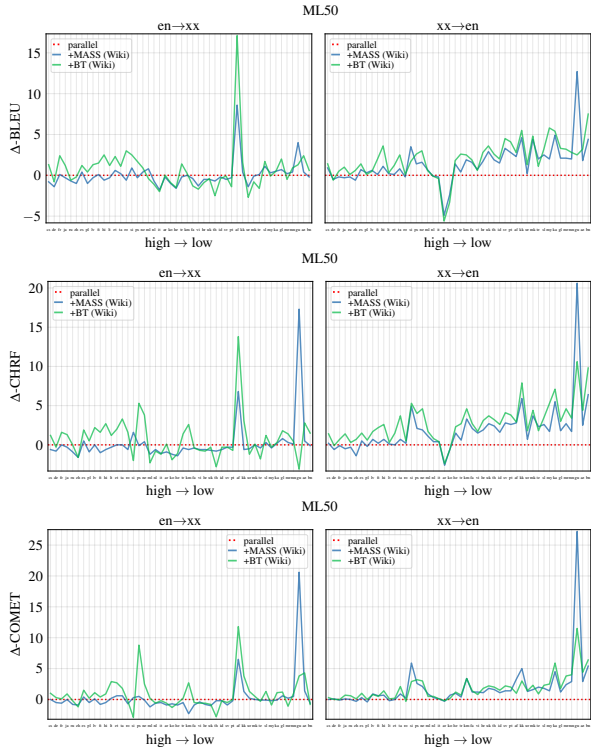
Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
parallel	53.3	45.5	46.8	61.0	52.4	45.4	50.6
Wiki							
+BART	52.8	46.2	47.9	61.0	54.6	53.3	52.6
+MASS	52.8	44.8	48.0	61.0	55.2	53.3	52.5
+BT	**55.4**	**49.7**	48.6	64.2	60.7	57.0	55.8
Mix							
+BART	51.7	44.5	47.6	60.8	54.8	52.8	52.1
+MASS	51.9	45.4	47.6	61.0	54.9	54.2	52.5
+BT	54.4	47.4	50.0	64.2	60.8	**59.0**	56.0
Mix+bal							
+BART	52.1	44.6	47.9	60.4	54.6	53.2	52.2
+MASS	52.5	45.3	47.9	60.6	55.6	54.9	52.9
+BT	55.2	48.8	**50.5**	**64.5**	**61.0**	58.9	**56.5**

(b) chrF scores (↑)

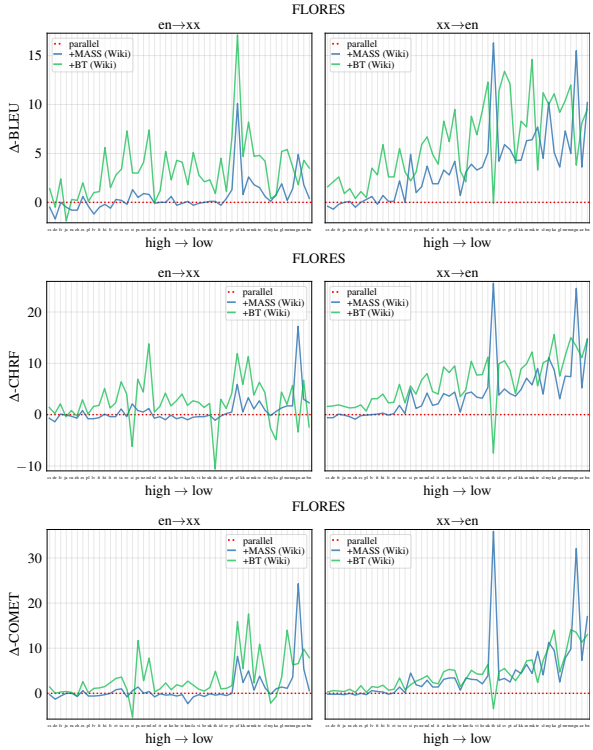
Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
parallel	80.3	76.4	69.9	83.4	79.2	73.1	76.7
Wiki							
+BART	79.8	76.6	70.9	83.6	81.2	80.2	78.5
+MASS	79.9	75.6	70.9	83.6	81.4	80.5	78.5
+BT	81.1	**80.2**	75.8	84.7	83.0	80.5	80.7
Mix							
+BART	78.9	75.3	70.5	83.4	81.2	80.0	78.1
+MASS	79.2	76.3	70.8	83.6	81.3	81.0	78.5
+BT	81.3	79.3	76.9	84.9	83.5	82.2	81.2
Mix+bal							
+BART	79.3	75.4	70.6	83.1	81.0	80.2	78.1
+MASS	79.5	76.3	70.7	83.4	81.8	81.5	78.7
+BT	**81.6**	80.1	**77.3**	**85.1**	**83.6**	**82.4**	**81.6**

(c) COMET scores (↑)

Table 7: Results of the Transformer-Big models on the **TICO-19** (Medical) test set and grouped by the monolingual split that has been used for training BT and DAE. Cells in red indicate worse scores than the baseline.

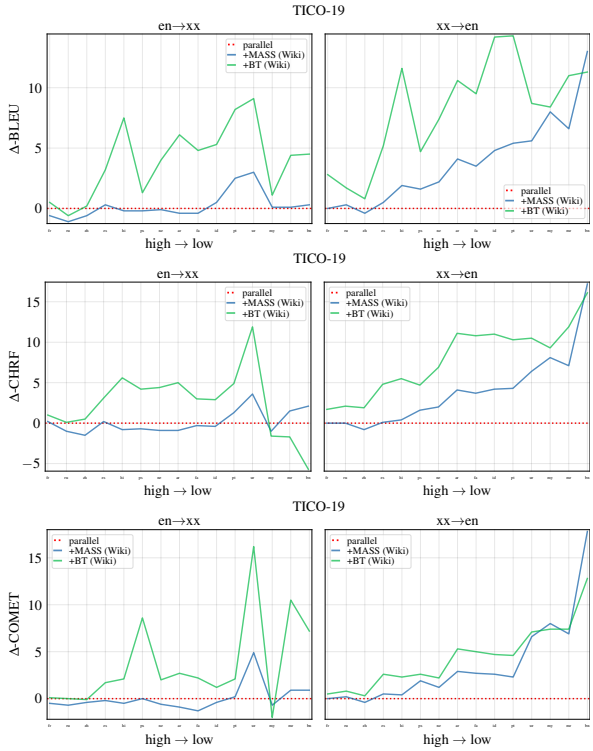


(a) Results on ML50 test sets.



(b) Results on FLORES (wiki) test sets.

(c) Results on NTREX (news) test sets.



(d) Results on TICO-19 (medical) test sets.

Table 8: Score (BLEU, chrF, COMET) differences between each model and the parallel-only baseline (red dotted line) across test sets, for models with the Transformer-Big architecture (370M).

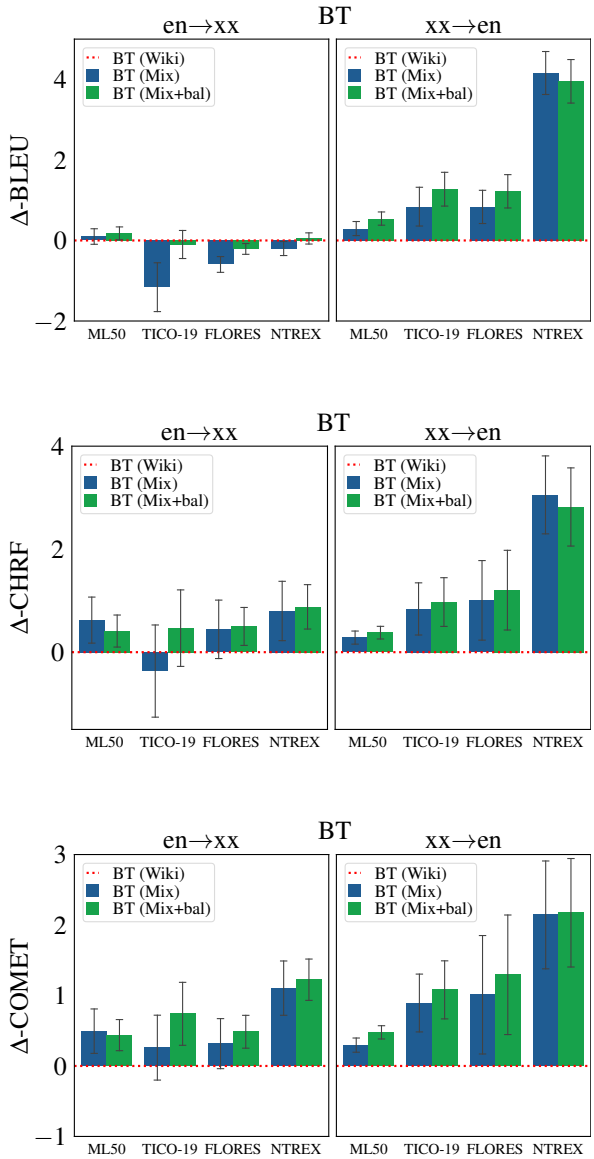


Figure 9: Score differences (Δ-X) of the **BT models** trained with the mixed-domain split with respect to the single-domain monolingual data (dotted red line). The top plot shows the Δ-BLEU scores, whereas the bottom shows the Δ-ChrF scores. To plot the bars, we use the mean Δ-X and the standard error.

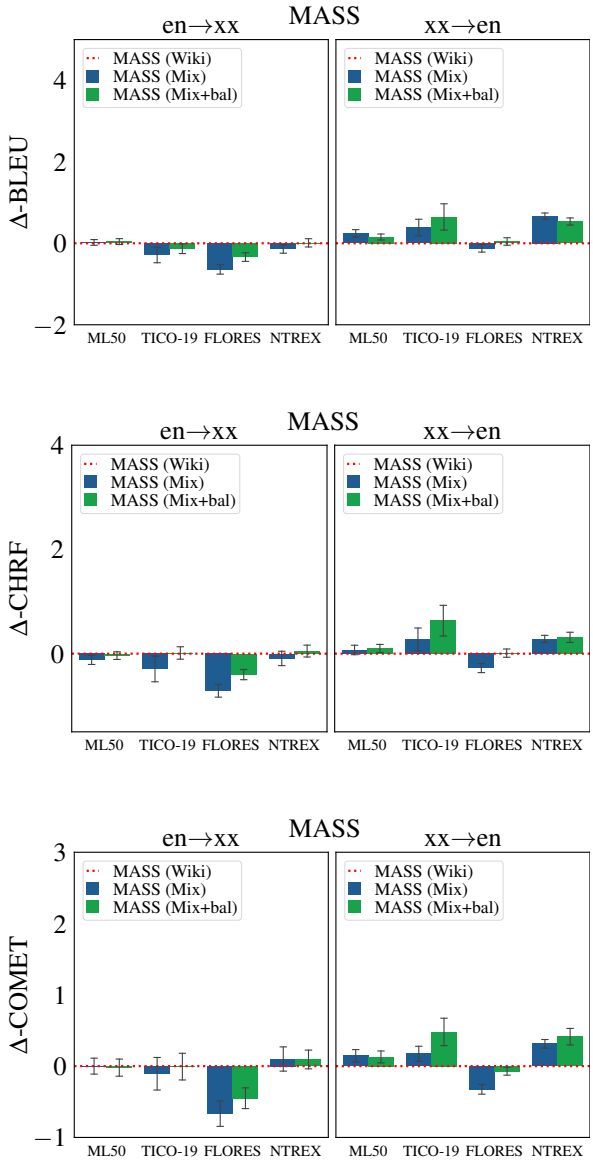


Figure 10: Score differences (Δ-X) of the **MASS (DAE) models** trained with the mixed-domain split with respect to the single-domain monolingual data (dotted red line). The top plot shows the Δ-BLEU scores, whereas the bottom shows the Δ-ChrF scores. To plot the bars, we use the mean Δ-X and the standard error.

B.2 Denoising Autoencoding Objectives

In this section, we extend the comparison of the two DAE objectives that is presented in Section 5.3 by including the results across all metrics and monolingual splits. Specifically, Table 9 shows the results with the balanced mixed-domain monolingual split, Table 10 with the unbalanced mixed-domain monolingual split, and Table 11 with the single-domain (Wikipedia) monolingual split. We observe that the differences are very small between models, but MASS outperforms BART by a small margin in most settings, similar to what is discussed in the main paper.

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
FLORES							
+BART	23.6	15.2	14.3	27.8	24.4	22.0	20.8
+MASS	**23.8**	**15.3**	**14.5**	**28.4**	**25.0**	**23.4**	**21.3**
NTREX							
+BART	21.8	13.1	**13.7**	25.0	23.2	21.0	19.4
+MASS	**21.9**	**13.3**	**13.7**	**25.7**	**23.8**	**22.1**	**19.8**
ML50							
+BART	**22.1**	**16.8**	21.3	26.8	26.1	28.1	23.5
+MASS	**22.1**	**16.8**	**21.5**	**27.2**	**26.6**	**28.8**	**23.8**
TICO-19							
+BART	31.2	14.0	15.1	31.8	26.0	24.1	23.7
+MASS	**31.5**	**14.4**	**15.2**	**32.6**	**27.2**	**26.4**	**24.5**

(a) BLEU scores (↑)

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
FLORES							
+BART	49.4	45.3	46.5	56.4	52.9	50.4	49.9
+MASS	**49.7**	**45.8**	**46.8**	**56.8**	**53.3**	**51.6**	**50.4**
NTREX							
+BART	47.4	42.9	44.1	54.4	51.8	48.2	47.9
+MASS	**47.6**	**43.3**	**44.3**	**54.6**	**52.1**	**49.3**	**48.3**
ML50							
+BART	47.9	44.1	47.4	55.1	50.4	50.6	49.1
+MASS	**48.0**	**44.4**	**47.6**	**55.2**	**50.6**	**51.2**	**49.3**
TICO-19							
+BART	52.1	44.6	**47.9**	60.4	54.6	53.2	52.2
+MASS	**52.5**	**45.3**	**47.9**	**60.6**	**55.6**	**54.9**	**52.9**

(b) chrF scores (↑)

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
FLORES							
+BART	81.9	79.5	73.4	83.3	81.0	78.5	79.3
+MASS	**82.3**	**80.0**	**74.0**	**83.7**	**81.6**	**79.7**	**79.9**
NTREX							
+BART	78.2	76.3	71.9	81.8	80.5	75.9	77.2
+MASS	**78.6**	**76.8**	**72.3**	**82.2**	**81.0**	**77.2**	**77.8**
ML50							
+BART	80.2	79.5	78.3	80.8	78.9	79.0	79.4
+MASS	**80.6**	**79.7**	**78.8**	**81.2**	**79.3**	**79.8**	**79.8**
TICO-19							
+BART	79.3	75.4	70.6	83.1	81.0	80.2	78.1
+MASS	**79.5**	**76.3**	**70.7**	**83.4**	**81.8**	**81.5**	**78.7**

(c) COMET scores (↑)

Table 9: Comparison of the DAE objectives with models trained on the **balanced mixed-domain**.

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
FLORES							
+BART	23.4	15.0	**14.2**	27.9	24.7	22.6	20.9
+MASS	**23.5**	**15.2**	14.1	**28.5**	**24.8**	**23.0**	**21.1**
NTREX							
+BART	21.6	13.0	13.6	25.4	23.4	21.5	19.5
+MASS	**21.7**	**13.1**	**13.7**	**26.0**	**23.9**	**22.1**	**19.8**
ML50							
+BART	21.8	16.7	21.4	27.1	26.3	28.4	23.6
+MASS	**22.0**	**16.8**	**21.5**	**27.4**	**26.5**	**28.9**	**23.8**
TICO-19							
+BART	30.5	13.9	14.9	32.5	26.6	24.2	23.7
+MASS	**31.1**	**14.3**	**15.2**	**33.0**	**26.9**	**25.6**	**24.3**

(a) BLEU scores (↑)

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
FLORES							
+BART	**49.5**	45.1	**46.4**	56.5	52.9	50.6	49.9
+MASS	49.4	**45.7**	**46.4**	**56.7**	**53.0**	**51.2**	**50.1**
NTREX							
+BART	47.4	42.8	43.9	54.5	51.7	48.4	47.9
+MASS	**47.5**	**43.2**	**44.1**	**54.7**	**52.0**	**49.1**	**48.2**
ML50							
+BART	47.7	44.0	47.4	**55.3**	**50.4**	50.8	49.1
+MASS	**47.9**	**44.4**	**47.5**	**55.3**	**50.4**	**51.2**	**49.3**
TICO-19							
+BART	51.7	44.5	**47.6**	60.8	54.8	52.8	52.1
+MASS	**51.9**	**45.4**	**47.6**	**61.0**	**54.9**	**54.2**	**52.5**

(b) chrF scores (↑)

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
FLORES							
+BART	81.8	79.1	73.5	83.4	81.1	78.6	79.3
+MASS	**82.0**	**80.0**	**73.7**	**83.7**	**81.3**	**79.3**	**79.7**
NTREX							
+BART	78.0	76.0	72.0	81.9	80.5	76.2	77.2
+MASS	**78.5**	**76.9**	**72.3**	**82.3**	**81.0**	**76.9**	**77.8**
ML50							
+BART	80.0	79.2	78.7	80.9	79.0	79.3	79.4
+MASS	**80.5**	**79.9**	**78.8**	**81.2**	**79.3**	**79.8**	**79.8**
TICO-19							
+BART	78.9	75.3	70.5	83.4	81.2	80.0	78.1
+MASS	**79.2**	**76.3**	**70.8**	**83.6**	**81.3**	**81.0**	**78.5**

(c) COMET scores (↑)

Table 10: Comparison of the DAE objectives with models trained on the **unbalanced mixed-domain**.

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
FLORES							
+BART	24.0	**15.6**	14.7	28.3	24.9	22.5	21.2
+MASS	**24.3**	15.5	**14.9**	**28.6**	**25.2**	**23.0**	**21.5**
NTREX							
+BART	21.9	**13.2**	13.4	**25.5**	**23.3**	20.8	19.4
+MASS	**22.1**	**13.2**	**13.8**	**25.5**	**23.3**	**21.2**	**19.5**
ML50							
+BART	22.0	**16.9**	**21.3**	27.0	**26.6**	27.9	23.6
+MASS	**22.1**	**16.9**	**21.3**	**27.1**	26.5	**28.5**	**23.7**
TICO-19							
+BART	**31.9**	**14.9**	15.1	**32.9**	26.3	24.2	24.2
+MASS	**31.9**	14.0	**15.4**	**32.9**	**27.0**	**24.6**	**24.3**

(a) BLEU scores (↑)

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
FLORES							
+BART	49.9	**46.2**	47.1	56.9	53.1	50.6	50.4
+MASS	**50.2**	46.1	**47.3**	**57.0**	**53.4**	**51.3**	**50.6**
NTREX							
+BART	47.6	**43.2**	44.0	54.5	51.7	47.9	47.9
+MASS	**47.9**	43.0	**44.2**	**54.6**	**51.9**	**48.5**	**48.1**
ML50							
+BART	47.9	**44.6**	47.4	55.1	50.5	50.4	49.1
+MASS	**48.1**	44.5	**47.5**	**55.2**	**50.6**	**50.9**	**49.3**
TICO-19							
+BART	**52.8**	**46.2**	47.9	**61.0**	54.6	**53.3**	**52.6**
+MASS	**52.8**	44.8	**48.0**	**61.0**	**55.2**	**53.3**	52.5

(b) chrF scores (↑)

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
FLORES							
+BART	82.5	**80.5**	74.4	**83.9**	81.5	78.9	80.0
+MASS	**82.8**	80.3	**74.6**	**83.9**	**81.7**	**79.6**	**80.2**
NTREX							
+BART	78.3	**76.7**	72.0	82.1	80.6	75.5	77.3
+MASS	**78.8**	76.4	**72.3**	**82.3**	**80.8**	**76.3**	**77.6**
ML50							
+BART	80.4	**80.2**	78.4	80.9	79.1	78.9	79.5
+MASS	**80.7**	79.9	**78.6**	**81.0**	**79.4**	**79.4**	**79.7**
TICO-19							
+BART	79.8	**76.6**	**70.9**	**83.6**	81.2	80.2	**78.5**
+MASS	**79.9**	75.6	**70.9**	**83.6**	**81.4**	**80.5**	**78.5**

(c) COMET scores (↑)

Table 11: Comparison of the DAE objectives with models trained on the **(Wikipedia) single-domain**.

B.3 Scaling

In this section, we report all of our results for the model scale analysis (§5.4). Tables 12, 13, 14, 15 show the results on the ML50, FLORES, NTREX and TICO19 test sets, respectively. For each test set, we report side-by-side the results from each evaluation metric.

Model Averages per Scale As it is not easy to extract meaningful patterns from the results in table format, we also plot the corresponding line plots with the average score of each method per model scale across metrics, in Figure 11 (BLEU), Figure 12 (chrF), and Figure 13 (COMET). We observe that the trends are overall the same across both metrics. All metrics agree that at small scales, MASS fails to outperform the baseline but becomes much more effective, compared to the baseline, as the scale increases. This further supports the findings discussed in the main paper.

However, we discover that metrics disagree with each other about the *degree* that scale benefits DAE/MASS. Specifically, we see that according to BLEU, DAE at the 1.6B scale is competitive with BT only on the ML50 test set, whereas chrF (middle column) and COMET (right column) suggest that DAE becomes much stronger with scale. In particular, according to COMET, at the 1.6B scale, MASS matches or outperforms BT on most test sets.

Model Averages per Resource-Level For completeness, we also include the plots with the scaling patterns of each model across resource levels and translation directions, in Figure 14 (BLEU; left column), Figure 15 (chrF; middle column), Figure 16 (COMET; right column). Overall, the results are consistent across metrics and test sets and the discussion in the main paper still holds.

However, we do discover one interesting discrepancy, which potentially relates to the observations of the previous paragraph. Specifically, in the chrF plots we see that BT in en→xx low-resource settings (bottom-left plot per test set) tends to become less effective than the parallel baseline in all test sets except for ML50. Recall that ML50 is the most distant test set with respect to the (Wikipedia) monolingual data. We do not have a reliable explanation for this observation.

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
Base							
parallel	18.8	14.8	**18.3**	23.7	**23.2**	24.9	20.6
+MASS	18.0	13.9	17.4	23.0	22.3	24.7	19.9
+BT	**19.7**	**14.9**	17.7	**24.3**	23.0	**25.5**	**20.8**
Big							
parallel	22.5	**17.3**	20.6	26.9	25.9	25.3	22.9
+MASS	22.1	16.9	21.3	27.1	26.5	28.5	23.7
+BT	**23.6**	**17.3**	**21.6**	**27.8**	**26.9**	**28.9**	**24.3**
XL							
parallel	25.2	18.4	21.4	30.1	29.7	28.4	25.4
+MASS	25.7	19.0	23.1	**31.1**	**30.8**	**31.1**	**26.7**
+BT	**26.5**	**19.3**	**23.2**	31.0	30.6	30.7	**26.7**

(a) BLEU scores (↑)

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
Base							
parallel	44.9	41.3	43.3	52.3	47.4	47.0	45.8
+MASS	43.9	39.9	42.1	51.3	46.3	46.9	44.9
+BT	**46.0**	**42.3**	**43.4**	**53.2**	**48.1**	**48.6**	**46.7**
Big							
parallel	48.6	45.0	46.3	55.3	49.3	47.0	48.3
+MASS	48.1	44.5	**47.5**	55.2	50.6	50.9	49.3
+BT	**49.7**	**45.4**	47.2	**56.4**	**51.6**	**51.7**	**50.1**
XL							
parallel	50.9	46.2	46.7	57.8	52.1	49.7	50.2
+MASS	51.2	46.9	**49.3**	58.6	53.4	52.9	51.8
+BT	**52.3**	**47.3**	48.2	**58.8**	**54.0**	**53.0**	**52.0**

(b) chrF scores (↑)

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
Base							
parallel	75.0	75.4	**73.1**	76.9	75.3	74.3	74.9
+MASS	73.2	73.2	71.9	76.0	74.4	74.8	73.9
+BT	**75.6**	**76.2**	72.9	**77.3**	**75.4**	**75.7**	**75.4**
Big							
parallel	80.9	80.5	77.0	80.8	78.0	75.6	78.6
+MASS	80.7	79.9	**78.6**	81.0	**79.4**	**79.4**	79.7
+BT	**81.8**	**80.9**	78.3	**81.3**	**79.4**	78.9	**80.0**
XL							
parallel	83.7	82.2	77.7	83.4	79.8	77.1	80.3
+MASS	84.0	82.9	**81.1**	**84.1**	81.1	**81.0**	**82.2**
+BT	**84.6**	**83.0**	79.6	83.8	**81.5**	80.1	81.9

(c) COMET scores (↑)

Table 12: Results of all methods across different model scales evaluated on the **ML50** (mixed-domain) test set. The BT and DAE models have used the (Wikipedia) single-domain monolingual split.

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
Base							
parallel	20.0	12.3	10.8	24.1	20.2	16.0	16.8
+MASS	19.0	11.7	11.3	23.5	19.6	17.9	16.8
+BT	**21.1**	**15.3**	**14.9**	**26.4**	**23.6**	**21.1**	**20.1**
Big							
parallel	24.8	15.3	13.2	28.6	22.4	16.0	19.4
+MASS	24.3	15.5	14.9	28.6	25.2	23.0	21.5
+BT	**26.0**	**18.8**	**17.7**	**30.8**	**28.4**	**24.9**	**24.1**
XL							
parallel	27.6	17.1	13.8	32.3	23.1	17.0	21.0
+MASS	28.3	18.4	16.2	33.5	25.6	23.5	23.7
+BT	**29.6**	**21.1**	**19.2**	**34.4**	**32.3**	**27.5**	**26.9**

(a) BLEU scores (↑)

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
Base							
parallel	46.2	41.1	40.9	53.5	48.8	42.6	45.1
+MASS	45.1	39.4	40.2	52.6	48.0	45.9	44.9
+BT	**47.9**	**45.2**	**44.5**	**55.8**	**53.0**	**48.8**	**48.9**
Big							
parallel	50.6	46.1	45.0	57.1	50.4	42.6	48.2
+MASS	50.2	46.1	47.3	57.0	53.4	51.3	50.6
+BT	**52.1**	**49.3**	**47.6**	**59.2**	**57.1**	**52.2**	**52.7**
XL							
parallel	53.0	47.9	45.6	60.0	51.4	44.9	49.9
+MASS	53.5	49.5	**49.5**	60.8	54.1	52.6	53.0
+BT	**54.8**	**51.5**	47.2	**62.0**	**60.2**	**54.4**	**54.6**

(b) chrF scores (↑)

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
Base							
parallel	76.9	73.1	64.9	79.7	75.7	67.1	72.4
+MASS	74.9	70.0	64.2	78.9	75.2	72.1	72.1
+BT	**78.4**	**76.9**	**71.4**	**81.1**	**78.3**	**72.8**	**76.1**
Big							
parallel	83.0	80.5	71.4	84.0	79.3	69.9	77.4
+MASS	82.8	80.3	74.6	83.9	81.7	**79.6**	80.2
+BT	**84.1**	**82.7**	**77.5**	**84.8**	**82.9**	77.2	**81.2**
XL							
parallel	85.6	83.1	72.7	86.3	81.0	73.2	79.7
+MASS	86.1	84.4	78.1	**86.9**	83.4	**82.5**	**83.3**
+BT	**86.7**	**84.8**	**78.4**	**86.9**	**85.6**	79.4	**83.3**

(c) COMET scores (↑)

Table 13: Results of all methods across different model scales evaluated on the **FLORES** (Wikipedia) test set. The BT and DAE models have used the (Wikipedia) single-domain monolingual split.

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
Base							
parallel	18.5	10.7	10.5	21.5	18.7	15.0	15.5
+MASS	17.8	10.2	10.8	20.9	18.2	16.7	15.5
+BT	**19.5**	**12.7**	**13.6**	**23.2**	**20.9**	**18.6**	**17.9**
Big							
parallel	22.4	13.2	12.4	25.1	21.1	15.1	17.8
+MASS	22.1	13.2	13.8	25.5	23.3	21.2	19.5
+BT	**23.3**	**15.5**	**16.0**	**27.4**	**25.1**	**21.8**	**21.2**
XL							
parallel	24.6	14.4	13.0	29.2	22.2	16.1	19.4
+MASS	25.1	15.6	15.0	29.9	24.5	21.7	21.6
+BT	**26.0**	**17.5**	**17.0**	**31.1**	**28.8**	**24.3**	**23.8**

(a) BLEU scores (↑)

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
Base							
parallel	44.4	38.9	38.5	51.6	47.5	40.6	43.2
+MASS	43.5	37.4	37.8	50.7	46.8	43.6	43.0
+BT	**45.7**	**41.9**	**41.5**	**53.5**	**50.9**	**45.3**	**46.2**
Big							
parallel	48.3	43.3	42.2	54.5	49.3	40.8	46.0
+MASS	47.9	43.0	**44.2**	54.6	51.9	**48.5**	48.1
+BT	**49.2**	**45.7**	**44.2**	**56.7**	**54.6**	48.3	**49.5**
XL							
parallel	50.3	44.7	42.8	57.5	50.7	43.0	47.7
+MASS	50.6	46.1	**46.4**	58.0	52.9	49.8	50.4
+BT	**51.7**	**47.7**	43.1	**59.4**	**57.5**	**50.2**	**51.2**

(b) chrF scores (↑)

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
Base							
parallel	72.6	69.8	63.2	78.2	75.1	65.0	70.2
+MASS	70.8	67.0	62.6	77.3	74.6	**69.5**	70.0
+BT	**73.2**	**72.4**	**68.3**	**79.2**	**77.2**	**69.5**	**73.0**
Big							
parallel	79.0	76.9	69.4	82.1	78.7	67.6	75.2
+MASS	78.8	76.4	72.3	82.3	80.8	**76.3**	77.6
+BT	**79.7**	**78.7**	**74.4**	**83.0**	**81.6**	73.4	**78.2**
XL							
parallel	81.9	79.4	71.0	84.5	80.4	70.2	77.4
+MASS	82.4	81.0	**76.2**	85.1	82.6	**79.0**	**80.8**
+BT	**83.0**	**81.4**	75.4	**85.2**	**84.3**	75.2	80.4

(c) COMET scores (↑)

Table 14: Results of all methods across different model scales evaluated on the **NTREX** (News) test set. The BT and DAE models have used the (Wikipedia) single-domain monolingual split.

Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
Base							
parallel	27.5	11.6	12.4	28.4	21.6	17.3	19.7
+MASS	26.7	11.1	12.5	27.5	21.3	20.0	19.9
+BT	**30.2**	**14.9**	**17.2**	**32.3**	**26.5**	**24.0**	**24.2**
Big							
parallel	32.3	14.3	14.4	32.4	24.2	17.4	22.3
+MASS	31.9	14.0	15.4	32.9	27.0	24.6	24.3
+BT	**34.5**	**18.4**	**19.8**	**36.8**	**32.2**	**28.7**	**28.3**
XL							
parallel	34.8	14.7	14.9	36.8	25.9	18.7	24.1
+MASS	35.2	16.1	16.1	38.0	28.1	25.2	26.4
+BT	**38.0**	**20.7**	**21.5**	**41.0**	**36.8**	**32.9**	**31.7**

(a) BLEU scores (↑)

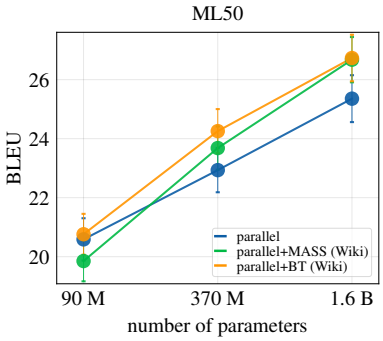
Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
Base							
parallel	49.4	41.2	42.9	57.8	50.2	45.1	47.7
+MASS	48.6	39.7	41.4	56.6	49.5	48.1	47.3
+BT	**52.1**	**44.8**	**46.1**	**60.7**	**55.7**	**52.8**	**52.0**
Big							
parallel	53.3	45.5	46.8	61.0	52.4	45.4	50.6
+MASS	52.8	44.8	48.0	61.0	55.2	53.3	52.5
+BT	**55.4**	**49.7**	**48.6**	**64.2**	**60.7**	**57.0**	**55.8**
XL							
parallel	55.1	45.5	47.4	64.2	53.7	47.2	52.1
+MASS	55.1	47.2	**49.5**	64.7	56.1	55.2	54.7
+BT	**58.0**	**50.7**	46.7	**67.0**	**64.1**	**60.5**	**57.6**

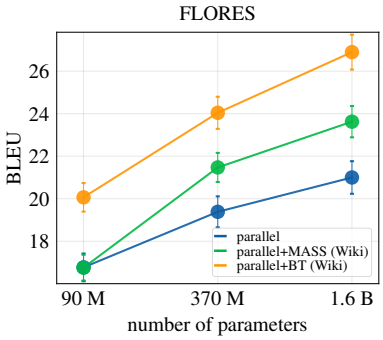
(b) chrF scores (↑)

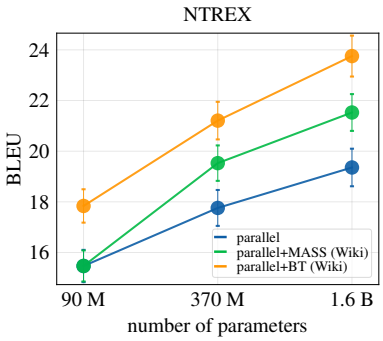
Model	en→xx			xx→en			Mean
	High	Med	Low	High	Med	Low	
Base							
parallel	75.6	70.8	65.7	80.5	76.0	70.5	72.9
+MASS	74.0	68.8	64.1	79.7	75.6	74.2	72.6
+BT	**77.0**	**75.0**	**71.0**	**82.2**	**78.6**	**76.0**	**76.4**
Big							
parallel	80.3	76.4	69.9	83.4	79.2	73.1	76.7
+MASS	79.9	75.6	70.9	83.6	81.4	**80.5**	78.5
+BT	**81.1**	**80.2**	**75.8**	**84.7**	**83.0**	**80.5**	**80.7**
XL							
parallel	82.4	76.8	70.4	85.5	80.9	75.7	78.3
+MASS	82.7	78.3	72.7	85.9	83.2	83.2	80.8
+BT	**83.2**	**81.3**	**75.6**	**86.2**	**85.4**	**83.7**	**82.4**

(c) COMET scores (↑)

Table 15: Results of all methods across different model scales evaluated on the **TICO-19** (Medical) test set. The BT and DAE models have used the (Wikipedia) single-domain monolingual split.







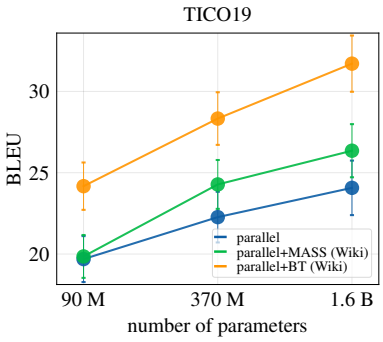
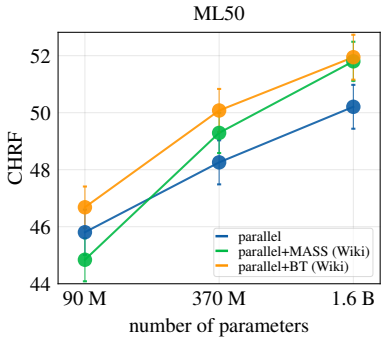
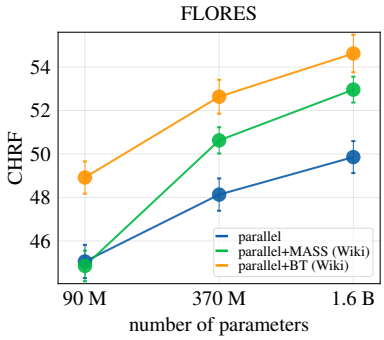
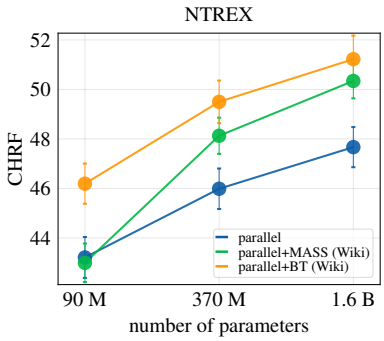


Figure 11: Average BLEU scores across model scales.







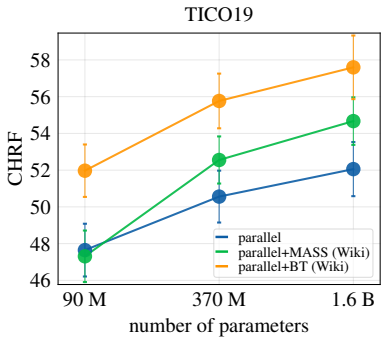
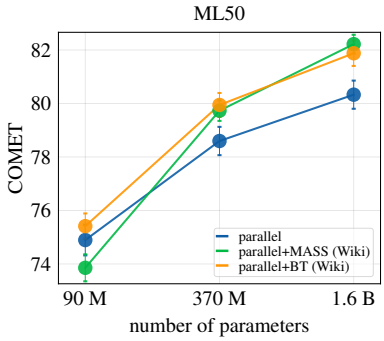
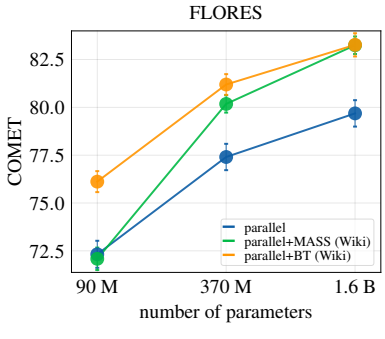
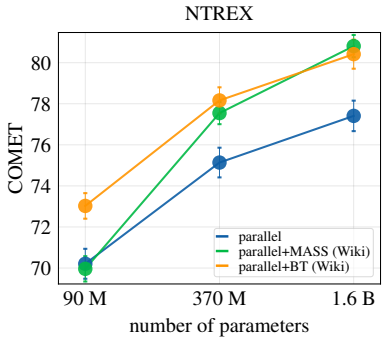


Figure 12: Average ChrF scores across model scales.







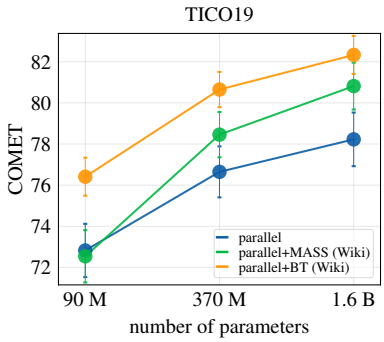


Figure 13: Average COMET scores across model scales.

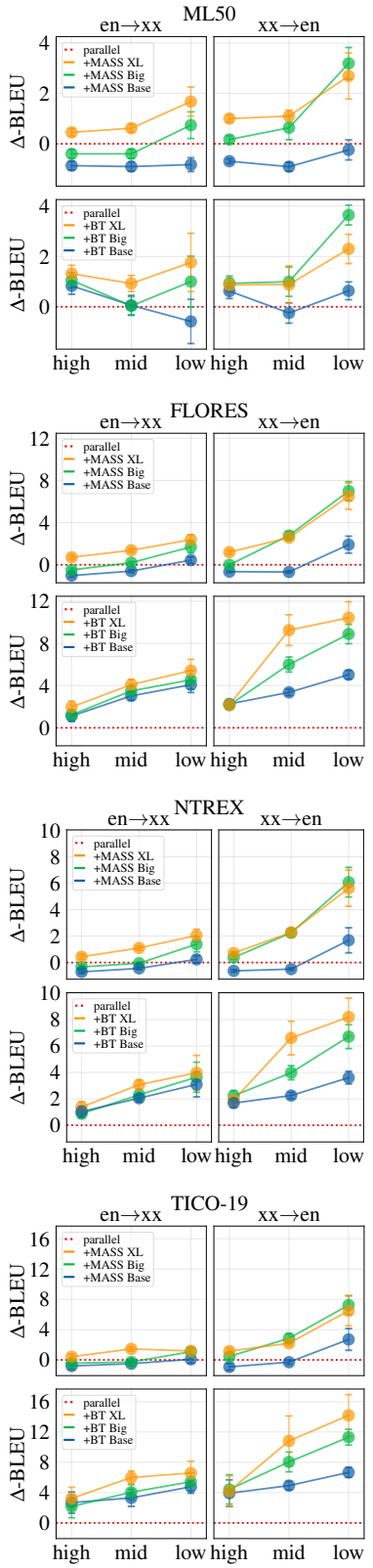


Figure 14: Mean BLEU differences (and standard error of the mean) per model with respect to the parallel-only baseline in the same scale (red dotted line).

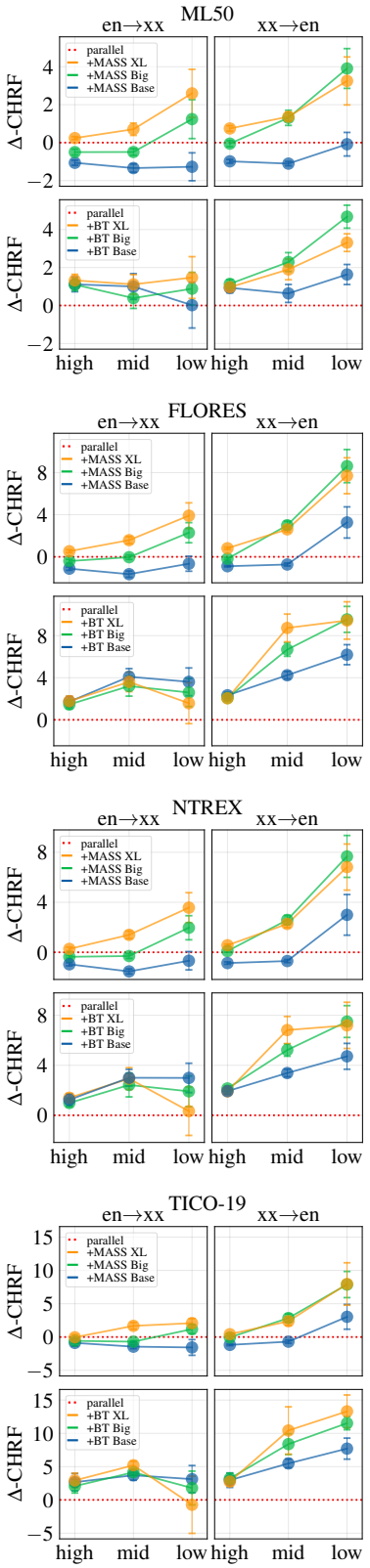


Figure 15: Mean chrF differences (and standard error of the mean) per model with respect to the parallel-only baseline in the same scale (red dotted line).

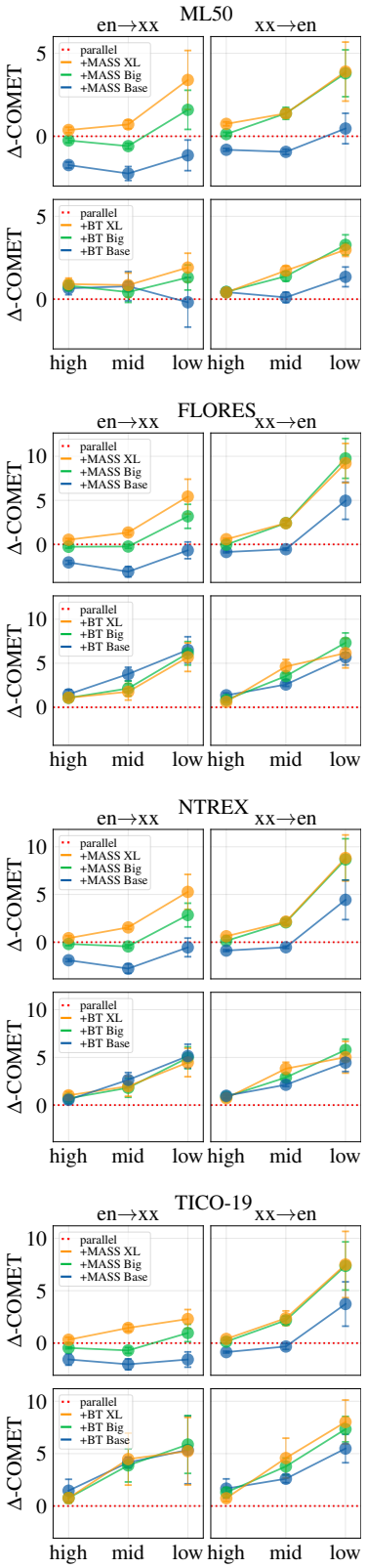


Figure 16: Mean COMET differences (and standard error of the mean) per model with respect to the parallel-only baseline in the same scale (red dotted line).

C Additional Tables and Figures

Group	Lang.	Parallel	Parallel + cap (10M)	wiki + cap (10M)	cc100 + cap (10M)	news + cap (10M)
high	cs	51,517,074	10,000,000	5,000,000	5,000,000	5,000,000
	de	45,992,835	10,000,000	5,000,000	5,000,000	5,000,000
	fr	38,507,539	10,000,000	5,000,000	5,000,000	5,000,000
	ja	17,203,227	10,000,000	5,000,000	5,000,000	5,000,000
	ru	13,599,766	10,000,000	5,000,000	5,000,000	5,000,000
	zh	11,173,646	10,000,000	5,000,000	5,000,000	5,000,000
	es	10,531,168	10,000,000	5,000,000	5,000,000	5,000,000
	pl	10,312,571	10,000,000	169,333	5,000,000	5,000,000
	lv	2,468,386	2,468,386	1,261,660	5,000,000	5,000,000
	fi	2,441,863	2,441,863	1,153,179	5,000,000	5,000,000
	hi	1,450,114	1,450,114	1,856,414	5,000,000	5,000,000
	lt	1,402,892	1,402,892	1,947,248	5,000,000	5,000,000
	iu	1,109,076	1,109,076	**1,892*	0	0
	et	1,064,974	1,064,974	2,585,642	5,000,000	5,000,000
medium	ta	612,747	612,747	2,119,411	5,000,000	2,861,282
	ro	600,019	600,019	3,604,671	5,000,000	5,000,000
	si	594,438	594,438	443,711	5,000,000	0
	ps	573,218	573,218	391,604	2,000,879	1,096,628
	ne	504,085	504,085	328,219	5,000,000	0
	ml	343,668	343,668	1,481,937	5,000,000	1,423,835
	nl	232,038	232,038	5,000,000	5,000,000	2,967,745
	it	226,385	226,385	5,000,000	5,000,000	5,000,000
	ar	225,678	225,678	5,000,000	5,000,000	5,000,000
	ko	223,750	223,750	5,000,000	5,000,000	5,000,000
	he	204,468	204,468	5,000,000	5,000,000	0
	tr	203,702	203,702	5,000,000	5,000,000	5,000,000
	km	183,934	183,934	256,007	3,398,559	0
	fa	142,128	142,128	5,000,000	5,000,000	5,000,000
	vi	127,117	127,117	5,000,000	5,000,000	0
	hr	116,866	116,866	2,556,084	5,000,000	5,000,000
	uk	104,021	104,021	5,000,000	5,000,000	2,222,071
low	th	91,245	91,245	514,270	5,000,000	0
	id	83,932	83,932	5,000,000	5,000,000	2,378,340
	sv	53,580	53,580	5,000,000	5,000,000	0
	pt	49,431	49,431	5,000,000	5,000,000	5,000,000
	af	41,268	41,268	1,260,811	5,000,000	428,151
	xh	37,900	37,900	**14,985*	437,761	0
	kk	27,618	27,618	1,674,930	5,000,000	3,869,280
	ur	25,188	25,188	1,133,339	5,000,000	0
	mk	24,022	24,022	1,953,775	5,000,000	863,917
	te	21,513	21,513	1,568,018	5,000,000	3,461,218
	sl	18,714	18,714	2,340,732	5,000,000	0
	my	17,980	17,980	943,634	1,229,875	0
	ka	12,292	12,292	264,710	5,000,000	0
	gl	9,491	9,491	2,358,124	5,000,000	0
	mr	9,203	9,203	644,383	5,000,000	827,586
	mn	7,145	7,145	332,251	5,000,000	0
	gu	6,535	6,535	340,779	4,767,339	3,042,472
	az	5,652	5,652	2,355,880	5,000,000	0
	bn	4,338	4,338	2,699,357	5,000,000	5,000,000

Table 16: The statistics of the parallel and training data we use for each language. The red-highlighted rows show the languages that we remove from our experiments.