Disentangling the Roles of Target-side Transfer and Regularization in Multilingual Machine Translation

Anonymous ACL submission

Abstract

Multilingual Machine Translation (MMT) benefits from knowledge transfer across different language pairs. However, improvements in one-to-many translation are only marginal compared to many-to-one translation. A widely held assumption is that knowledge transfer barely plays a role in the target-side of MMT. The observed improvements in one-to-many MT are instead attributed to two possible reasons: increasing the amounts of source language data and target language regularization. In this paper, we conduct a large-scale study that varies the target-side languages along two dimensions, i.e., linguistic similarity and corpus size, to show the interplay between different factors (knowledge transfer, source data size, language regularization) for improving one-to-many translation. First, we find that positive knowledge transfer *does* occur on the target-side, which greatly benefits low- and medium-resource language pairs. Moreover, the performance discrepancy across different target languages also shows that increasing the source-side data cannot be the main reason for improving one-to-many MT. Furthermore, we show language regularization plays a crucial role in benefiting translation performance by enhancing the generalization ability and model inference calibration. We find a simple but effective way to utilize distant target data with the aim of regularizing the model, which surprisingly leads to translation performance gains.

1 Introduction

Multilingual Machine Translation (MMT) enables a single model to translate among multiple language pairs by joint training (Dong et al., 2015; Johnson et al., 2017). The improvements in translation quality, especially for low-resource languages, are generally attributed to transfer learning (Zoph et al., 2016; Lakew et al., 2018; Kocmi and Bojar, 2018; Stap et al., 2023). However, MMT suffers from a performance gap where the gains in one-to-many translation are not as substantial as in many-to-one translation (Dabre et al., 2020; Tang et al., 2020; Yang et al., 2021; Chiang et al., 2021; Chowdhery et al., 2022). Empirical studies (Johnson et al., 2017; Aharoni et al., 2019) also show little or even no benefit in one-to-many translation compared to their bilingual baselines, leading to the hypothesis that positive transfer does not occur on the target-side (Arivazhagan et al., 2019).

The challenge of knowledge transfer in one-to-many translation is attributed to the inherent characteristics of translating into *distinct* target languages. The necessity of the target language-specific representations in the translation process hinders knowledge transfer as transfer learning prefers language-invariant representations (Kudugunta et al., 2019). On the other hand, Arivazhagan et al. (2019) and Aharoni et al. (2019) indicate that the increasing amounts of source language data and regularization induced by multiple target languages are possible reasons for the observed benefits in massively MMT scenarios.

Nevertheless, the extent to which positive knowledge transfer occurs on the target-side still remains unclear. Furthermore, a comprehensive analysis of the interplay between different factors, i.e., knowledge transfer, source data size, and regularization, in one-to-many translation is lacking. This hinders the optimization of MMT performance.

To understand the impact of knowledge transfer, we conduct comprehensive controlled experiments with varying target languages along two dimensions, i.e., linguistic similarity and corpus size. We select a set of bilingual out-of-English translation tasks, e.g., English to German, as main language pairs. Subsequently, we add different auxiliary target language pairs to the main language pairs, considering variations in auxiliary language families, written scripts, data sizes, and target language numbers. Our experimental results show a consistent positive correlation between the improvements and their translation task relatedness, i.e., increasing the amounts of similar target languages enhances positive knowledge transfer for the main language pair. These findings confirm the existence of knowledge transfer on the target-side and also clearly show factors that influence target-side transfer, i.e., target data size, number of translation tasks, and linguistic similarity. Meanwhile, the performance differences induced by various target languages also indicate that increasing source data is not the main reason for improving one-to-many MT.

Apart from knowledge transfer, we find that small amounts of distant auxiliary target data can act as an effective regularizer to yield improvements in translation quality. To understand why language regularization plays a role, we show it benefits translation performance by reducing generalization errors and improving inference calibration. With introducing small auxiliary target data, the translation model is implicitly calibrated so that the confidences of their predictions are more aligned with the accuracies of their predictions.

To summarize, we show how different factors i.e., knowledge transfer, source data size, and regularization, play roles in one-to-many translation. We first confirm the existence of positive knowledge transfer on the target-side, and show how linguistic similarity and data size mutually influence the extent of transfer learning in one-to-many translation. Meanwhile, we find that increasing source data plays a smaller role in improving one-to-many MT. Finally, our investigation of language regularization provides a simple yet effective way to boost machine translation performance by leveraging distant auxiliary data.

2 Background

In this section, we introduce the study of transfer learning, source data, and regularization in MMT.

2.1 Transfer Learning

Transfer learning is defined as improving a learner from one task by leveraging information from a related task (Weiss et al., 2016). An example is seen in MMT, where training models on multiple language pairs benefits resource-poor languages by leveraging shared linguistic information and parameters from other languages (Zoph et al., 2016; Murthy et al., 2019).

However, in the case of one-to-many machine translation, it leads to much more marginal gains than many-to-one translation. This performance discrepancy is caused by the challenges of target-side transfer. Aharoni et al. (2019) empirically emphasizes such difficulty of transfer on the target-side by showing the marginal benefits, even for low-resource language pairs, in a large-scale one-to-many translation. Dabre et al. (2020) indicate that the reason behind this challenge is mainly due to its characteristics of representations on the decoder side, where each target data has an independent output distribution and the decoder representations are more sensitive to the target languages (Kudugunta et al., 2019). Wang et al. (2018) further supports this claim by keeping target language-specific parameters to improve the one-to-many translation. This increases uncertainties on the effectiveness of transfer learning on the target-side, which oppositely prefers language-invariant representations.

Despite previous works (Gao et al., 2020; Shaham et al., 2022) indicating that linguistic similarity matters to encourage positive target-side transfer, their findings are limited to scenarios where knowledge is transferred from high-resource to low-resource. Fernandes et al. (2023) conversely shows that no impact of linguistic similarity on the translation performance for translating into two high-resource target languages, with an example of translating English into {French, Chinese} and English into {French, German}.

Overall, these studies show an inconsistent view towards the target-side transfer, particularly about whether this transfer exists and what factors influence it. This inconsistency indicates the importance of exploring target-side transfer in one-to-many MT and the impact of different factors on it.

2.2 Source Data Size

In English-centric one-to-many translation, the improvements in translation performance are attributed to the increasing source-side data instead of the target-side (Arivazhagan et al., 2019). The increasing source of English data results in better encoder representations to further benefit translation performance. However, it is still unclear whether the source data can be an entire reason to explain all the improvements.

2.3 Regularization

The multilingual training regime is known as a source of regularization, which improves the generalization ability of the models (Neubig and Hu, 2018; Aharoni et al., 2019; Dabre et al., 2020).

However, the effects of language regularization induced by multiple target tasks are under-explored, compared to other regularization techniques, such as dropout (Srivastava et al., 2014) and label smoothing (Szegedy et al., 2015). Dropout randomly selects activations to be "dropped out" during training. This randomness introduced by dropout encourages the network to learn robust and generalized representations (Liang et al., 2021). Another common regularization technique, label smoothing, regularizes the model by penalizing the output confidence. It has also been shown that these changes in output confidence introduced by label smoothing could implicitly enhance machine translation model calibration (Müller et al., 2019), thereby improving translation performance. In line with this, we aim to investigate language regularization in one-to-many translation to understand when and why it is effective.

3 Experimental setting

Model. We follow the setup of the Transformer base model (Vaswani et al., 2017). More details on model hyperparameters can be found in Appendix B.

Data. We choose three main language pairs in different language families and written scripts: English-into-German (En→De), English-into-Russian (En→Ru), and English-into-Spanish (En→Es). The training data for the main language pairs En→De, En→Ru, and En→Es are from WMT13, WMT14, and WMT22 respectively. To mimic low- and medium-resource settings, we randomly sample 100K and 1M translation pairs from each language pair respectively. To observe the impact on high-resource settings, we use the full training corpus for En→De (4.5M examples). For different controlled experiments, we cover 20 auxiliary target language pairs to train with the main translation tasks. We randomly sample the auxiliary covered language pairs from WMT and CCMatrix[1]. The detailed statistics of the main and auxiliary language pairs are shown in Appendix C.

Training and Evaluation. We use the Fairseq (Ott et al., 2019) toolkit to train transformer models. All models are trained with the Adam optimizer (Kingma and Ba, 2017) for up to 100K steps, with a learning rate of 5e-4 with an inverse square root scheduler. Dropout rate of 0.3 and label smoothing of 0.2 are used. Each model is trained on one A6000 GPU with a batch size of 25K tokens. We choose the best checkpoint according to the average validation loss of all language pairs. The data is tokenized with the SentencePiece tool (Kudo and Richardson, 2018) and we build a shared vocabulary of 32K tokens. We add language ID tokens to the vocabulary and prepend the language ID token to each source and target sequence to indicate the target language (Johnson et al., 2017). For evaluation, we employ beam search decoding with a beam size of 5. BLEU scores are computed used detokenized case-sensitive SacreBLEU[2] (Post, 2018).

4 Target-side Transfer

In this section, we aim to empirically reveal whether target-side transfer occurs in one-to-many machine translation. To achieve this, we select three main language pairs: En→De, En→Es, En→Ru, and train each main language pair with different auxiliary target languages to investigate the target-side transfer in multilingual machine translation for influencing main language pairs.

4.1 Changes in Target Language

Here, we introduce different auxiliary target languages with variations in linguistic similarity and data size. The varying auxiliary target data size represents the true distribution of varied data in multilingual machine translation.

4.1.1 Setup

For each main language pair (En→X), we train it with an auxiliary language pair (En→Y) that differs in language family and written script. In Appendix A, Table 4 presents the linguistic information about the main and auxiliary target languages. For the auxiliary target data training with the main low-resource language pair, we vary its data size with the proportion from 10% to 1000% of the main low-resource language pair. For the auxiliary target data training with the medium- and high-resource setting, we vary its data size with the proportion from 1% to 200% of the main language pair. To mitigate the variance in the quality of sampled auxiliary target language pairs, we run the experiment with three different randomly sampled sets.[3] Tables 1 and 2 show the averaged results of three

[1] https://opus.nlpl.eu/CCMatrix.php

[2] nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1

[3] We use one random sample set for high-resource auxiliary data due to computational constraints.

En→De (Baseline: 7.4)						**En→De (Baseline: 20.0)**					
$\alpha\%$	en→de	en→nl	en→et	en→ru	en→zh	$\alpha\%$	en→de	en→nl	en→et	en→ru	en→zh
10%	$8.5_{0.4}$	$7.9_{0.7}$	$8.2_{0.6}$	$8.6_{0.5}$	$8.9_{0.8}$	1%	$20.0_{0.4}$	$20.2_{0.4}$	$20.5_{0.2}$	$20.7_{0.3}$	$20.8_{0.5}$
50%	$10.2_{0.3}$	$10.3_{0.6}$	$10.5_{0.6}$	$10.9_{0.3}$	$11.5_{0.4}$	10%	$20.3_{0.2}$	$21.0_{0.3}$	$20.7_{0.4}$	$21.2_{0.6}$	$21.8_{0.6}$
100%	$11.6_{0.4}$	$11.3_{0.4}$	$10.9_{0.2}$	$11.0_{0.4}$	$12.1_{0.2}$	50%	$22.1_{0.4}$	$21.6_{0.5}$	$21.3_{0.1}$	$21.2_{0.2}$	$21.6_{0.2}$
500%	$15.9_{0.3}$	$14.0_{0.2}$	$13.7_{0.3}$	$13.4_{0.2}$	$13.5_{0.3}$	100%	$23.4_{0.2}$	$22.2_{0.2}$	$21.2_{0.2}$	$21.0_{0.2}$	$21.2_{0.2}$
1000%	$19.9_{0.1}$	$16.2_{0.2}$	$15.3_{0.1}$	$14.1_{0.2}$	$14.2_{0.1}$	200%	$24.5_{0.1}$	$22.2_{0.0}$	$20.2_{0.0}$	$20.0_{0.0}$	$20.7_{0.0}$

En→Ru (Baseline: 11.9)						**En→Ru (Baseline: 18.4)**					
$\alpha\%$	en→ru	en→uk	en→cs	en→de	en→zh	$\alpha\%$	en→ru	en→uk	en→cs	en→de	en→zh
10%	$12.0_{0.4}$	$11.8_{0.6}$	$11.6_{0.6}$	$11.7_{0.2}$	$12.0_{0.4}$	1%	$18.1_{0.3}$	$18.6_{0.5}$	$18.7_{0.8}$	$18.7_{0.5}$	$18.9_{0.2}$
50%	$12.8_{0.3}$	$13.0_{0.5}$	$12.2_{0.2}$	$12.4_{0.3}$	$12.6_{0.1}$	10%	$18.6_{0.5}$	$18.9_{0.2}$	$19.1_{0.1}$	$18.9_{0.2}$	$19.1_{0.3}$
100%	$14.0_{0.2}$	$13.3_{0.3}$	$12.6_{0.1}$	$12.7_{0.2}$	$12.8_{0.4}$	50%	$19.5_{0.2}$	$19.3_{0.3}$	$18.8_{0.1}$	$18.4_{0.2}$	$18.7_{0.1}$
500%	$15.7_{0.2}$	$14.7_{0.2}$	$14.2_{0.1}$	$14.4_{0.2}$	$14.6_{0.1}$	100%	$20.1_{0.1}$	$19.5_{0.2}$	$19.1_{0.1}$	$18.6_{0.2}$	$18.2_{0.1}$
1000%	$18.6_{0.3}$	$15.4_{0.1}$	$14.7_{0.2}$	$14.6_{0.2}$	$14.3_{0.2}$	200%	$22.4_{0.1}$	$20.5_{0.0}$	$18.5_{0.0}$	$17.2_{0.0}$	$17.1_{0.0}$

En→Es (Baseline: 16.9)						**En→Es (Baseline: 28.6)**					
$\alpha\%$	en→es	en→pt	en→nl	en→ru	en→zh	$\alpha\%$	en→es	en→pt	en→nl	en→ru	en→zh
10%	$17.1_{0.2}$	$17.0_{0.4}$	$17.3_{0.6}$	$17.2_{0.3}$	$17.6_{0.8}$	1%	$28.6_{0.3}$	$28.6_{0.1}$	$28.7_{0.2}$	$28.8_{0.2}$	$28.7_{0.5}$
50%	$19.0_{0.2}$	$18.1_{0.3}$	$18.5_{0.6}$	$19.0_{0.2}$	$19.5_{0.3}$	10%	$29.4_{0.2}$	$29.0_{0.3}$	$29.1_{0.2}$	$29.3_{0.4}$	$29.2_{0.3}$
100%	$20.9_{0.4}$	$19.1_{0.3}$	$19.4_{0.3}$	$19.1_{0.3}$	$21.0_{0.2}$	50%	$29.9_{0.4}$	$29.2_{0.5}$	$29.4_{0.2}$	$29.4_{0.2}$	$29.4_{0.1}$
500%	$27.1_{0.3}$	$23.2_{0.2}$	$21.5_{0.3}$	$22.8_{0.3}$	$23.0_{0.2}$	100%	$30.5_{0.3}$	$29.5_{0.3}$	$29.2_{0.1}$	$29.0_{0.3}$	$29.2_{0.4}$
1000%	$29.4_{0.2}$	$25.2_{0.4}$	$23.2_{0.1}$	$22.4_{0.3}$	$22.2_{0.1}$	200%	$31.8_{0.2}$	$29.6_{0.0}$	$28.9_{0.0}$	$28.3_{0.0}$	$28.0_{0.0}$

Table 1: BLEU scores (including variance) for three main tasks: En→De, En→Es, and En→Ru in low-resource 100K (left) and medium-resource 1M (right) settings when training with different auxiliary language pairs. $\alpha\%$ represents the auxiliary training data size. For low-resource setting, $\alpha\%$ ranges from 10% to 1000% of the proportion of the low-resource setting size. For medium-resource setting, $\alpha\%$ ranges from 1% to 200% of the proportion of the medium-resource setting size. The color block represents the extent of positive transfer, with the darker shades indicating a stronger positive transfer effect.

main translation tasks in low-, medium-and high-resource settings when training with different target languages, along with the corresponding variance.

4.1.2 Discussion

First, we show **positive knowledge transfer occurs on the target-side**, which benefits low-/medium-resource language pairs. This target-side positive transfer is highly correlated with translation task relatedness, i.e. linguistic similarity. Specifically, for low- and medium-resource settings (Table 1), increasing the amounts of similar target languages improves the positive knowledge transfer for the main language pairs, i.e. 9 BLEU points improvements for the low-resource En→De task when training with 1000% En→Nl. However, training with the same amounts of a distant target task cannot achieve similar improvements, such as En→Zh. The varying performance for the main tasks when training with different target-side languages shows that the increasing source English data (Arivazhagan et al., 2019) cannot be entirely confirmed as the sole reason for the improvements.

Second, we demonstrate that **negative transfer also exists with increasing amounts of target data.** For medium-resource settings, increasing the size of distant auxiliary languages gradually shows the negative transfer for main language pairs. For the high-resource setting (Table 2), negative transfer almost occurs in training with every auxiliary language pair. It still correlates with linguistic similarity where distant data results in more performance drops than similar ones. This is in line with (Wang et al., 2019) where they show that divergence between the joint distributions of tasks is the root of the negative transfer.

Third, we find that **the gains for low- or medium-resource tasks in one-to-many translation cannot be fully attributed to transfer learning.** The small amount of data can also improve the translation performance of main language pairs and do so without any correlation with linguistic similarity. In Table 1 (right), joint training with 10% distant language pairs can even lead to better translation performance for all main language tasks than using 10% similar data. 10% of En→Zh data can even lead to around 2 BLEU points improvement for the En→De task in a medium-resource setting. The gains resulting from the small size

	En→De (Baseline: 26.1)			
$\alpha\%$	en→nl	en→et	en→ru	en→zh
1%	26.6	26.3	26.0	26.0
10%	25.9	25.9	25.7	25.8
50%	25.9	25.0	25.2	24.8
100%	25.7	25.4	24.4	24.4
200%	25.2	24.8	23.4	23.1

Table 2: BLEU scores for the main language pair En→De in high-resource setting 4.5M. $\alpha\%$ ranges from 1% to 200% of the proportion of the high-resource setting size. The color block represents the extent of negative transfer, with the darker shades indicating a stronger negative transfer effect.

of distant auxiliary data show the role of language regularization. By joint training with auxiliary low-resource target tasks, uncertainties are increased for the model to prevent over-fitting on the main tasks. Further discussion is shown in Section 5.

4.2 Changes in Task Number

To further validate the previous findings, we expand the scenario from training with a single target task to incorporating multiple tasks. We control the total amount of auxiliary training data to ensure a fair comparison.

4.2.1 Setup

We train the main translation task En→De in different resource levels with an increasing number of auxiliary target language pairs from two groups (Table 5 in Appendix A): (1) Similar Group: the Germanic[4] language family with Latin scripts; (2) Distant group: the Slavic language family with Cyrillic scripts. The number of target language pairs is set as 1, 4, 8. The auxiliary target data size is evenly distributed among all target languages and controlled at 50% and 1000% for low-resource, and 10% and 200% for medium- and high-resource. Figure 1 shows the impact of task number when training with auxiliary tasks from different linguistic groups.

4.2.2 Discussion

We show that **increasing the task number has little impact on the target-side knowledge transfer**, since our findings are similar for two tasks (Section 4.1): (1) Positive transfer highly correlates with linguistic similarity when the auxiliary data size is large; (2) Small distant auxiliary target

[4]Due to data scarcity, we pick two target languages from the Romance language family, Galician, and Spanish. Romance and Germanic language families are close.

(a) Data size: 50% (b) Data size: 1000%

(c) Data size: 10% (d) Data size: 200%

(e) Data size: 10% (f) Data size: 200%

Figure 1: Translation quality for En→De for a low-resource 100K (above), medium-resource 1M (middle) and high-resource 4.5M (below) when training with different auxiliary task numbers and different linguistic groups. Data size represents the total amount of auxiliary target training data.

data can also benefit the low- and medium-resource main tasks, which is attributed to regularization. Interestingly, for the medium- and high-resource settings, increasing the auxiliary target task number from the large-size distant linguistic group (200%) can mitigate the negative transfer to some extent. One possible explanation for this is that the negative training signal from one distant language pair becomes weaker when increasing the task number in controlled data size. This result also corroborates similar findings (Shaham et al., 2022) where they find more than one unrelated language helps for the translation task with less data.

In summary, Section 4 shows the target-side transfer in one-to-many translation. Based on the empirical findings on main language pairs, we show that target-side transfer transfers positive knowledge. Linguistic similarity and target data size mutually play a role in it. Meanwhile, we show that the source data cannot be the sole reason for improving one-to-many translation due to the close correlation between translation performance and target data. Furthermore, we find that the small size of distant auxiliary target languages can also improve translation performance. These gains cannot be fully attributed to target-side transfer, and we indicate another important factor, i.e., regularization, which is discussed in the next section.

5 Language Regularization

The previous section shows low- and medium-resource translation tasks benefit from language regularization. In this section, we aim to further investigate the effectiveness of language regularization in one-to-many MT from two angles: generalization ability (Section 5.1) and model calibration (Section 5.2). In the end, we provide a simple but effective way to enhance the machine translation performance with the help of language regularization (Section 5.3).

5.1 Reducing Generalization Error

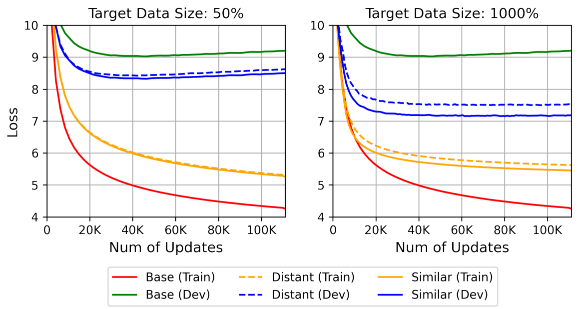
Reducing generalization errors is one of the benefits of regularization, which can be reflected by measuring the inconsistency between training and validation performance. Here, we show the regularization effects in one-to-many translation by comparing their learning curves for the training and valid losses.

5.1.1 Setup

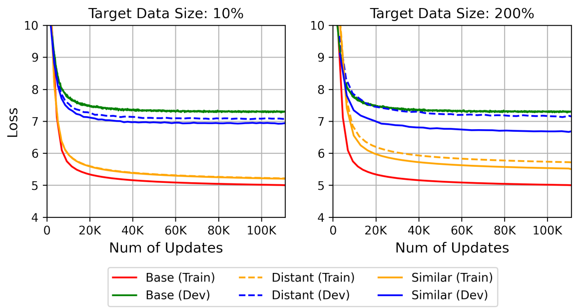
Different target languages have various levels of regularization effects. As we shown in Section 4.1, low- and medium-resource main language pairs benefit from regularization. Thus, we choose the multilingual models trained on low- and medium-resource En→De tasks with two linguistic groups shown in Section 4.2. For the low-resource En→De setting (100K), we select the auxiliary target data size to be 50% and 1000% of the low-resource size. For the medium-resource En→De setting (1M), we select the target data size to be 10% and 200% of the medium-resource size. Figure 2 shows the learning curves En→De under different multilingual training settings.

5.1.2 Discussion

First, **regularization induced by the small size of auxiliary target tasks can reduce the generalization errors** in one-to-many translation. Figure 2a shows that the baseline bilingual low-resource En→De model has a large gap between training and validation loss during training. This indicates that low-resource models can easily overfit and cannot generalize well to unseen data. Surprisingly, 50% of distant auxiliary data can reduce the validation loss for the main low-resource En→De task. This observation aligns with previous finding in Section 4.2 that distant auxiliary target languages benefit the main task performance. It confirms our hypothesis that regularization plays a crucial role in the gains via improving generalization ability.



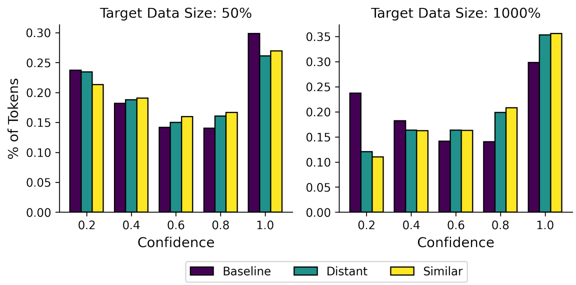
(a) En→De in Low-resource (100K)



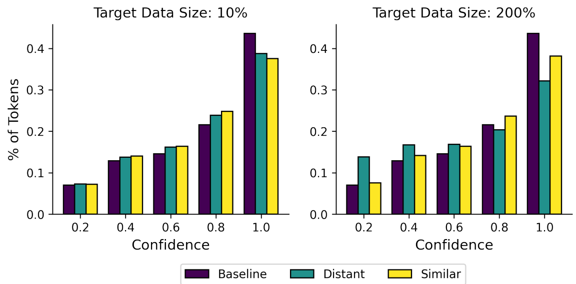
(b) En→De in Medium-resource (1M)

Figure 2: Loss curves for En→De translation tasks under low-resource 100K (a) and medium-resource 1M settings (b), with varying target linguistic groups (similar and distant) and varying auxiliary target data sizes.

Second, regularization effects from the large size of auxiliary target tasks can only reduce generalization errors for low-resource language pairs. Increasing the auxiliary target data size (+1000%) leads to better generalization ability for low-resource En→De, and the linguistically similar group shows slightly better effectiveness than the distant ones. This difference shows that positive target-side transfer also helps for better generalization ability since they exhibit a strong and transferrable training signal for the main low-resource task. The same holds for the medium-resource En→De setting (Figure 2b). However, when training with a large target data size (+200%), a distant linguistic group cannot further reduce the generalization errors.



(a) En→De in Low-resource (100K)



(b) En→De in Medium-resource (1M)

Figure 3: Confidence histograms for En→De translation tasks under low-resource (100K) (a) and mid-resource (1M) settings (b), with varying target linguistic groups (similar and distant) and total target data sizes.

This reflects that **the role of regularization is not always positive, heavily depending on the target linguistic similarity level and the data size.**

5.2 Improving Inference Calibration

Another benefit of regularization is to increase the model's uncertainty by penalizing output confidence, e.g., label smoothing. This regularization technique improves model calibration by making the confidence of its predictions more accurate for true accuracy (Müller et al., 2019). Wang et al. (2020) emphasizes the importance of calibrating confidence during inference for MT and regularization is a key factor. Motivated by these findings, we aim to investigate whether regularization induced by different target tasks has a similar impact on both output confidence and inference calibration.

In general, model calibration is measured by the expected calibration error (ECE) which calculates the difference in expectation between confidence and accuracy. As shown in Equation 5.2, ECE divides predictions into M bins $\{B_1, ..., B_M\}$ based on their confidence and calculates a weighted average of the bin's accuracy/confidence difference.[5]

[5] N is the number of prediction samples and $|B_m|$ is the number of samples in the m-th bin

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} |acc(B_m) - conf(B_m)| \quad (1)$$

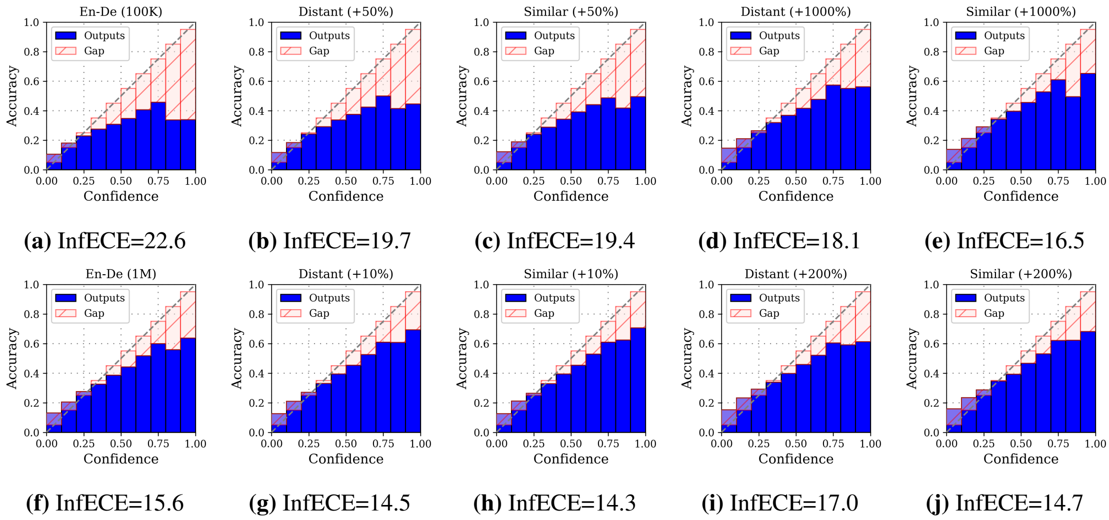
In MT, the prediction target token is $\hat{y} = argmax_{y \in V} P(y)$ and the confidence is $P(\hat{y})$. The accuracy denotes whether the prediction $\hat{y}$ is correct. However, calculating the prediction accuracy during inference is challenging because it requires building complex alignments between generated tokens and the ground truth. Wang et al. (2020) propose using the Translation Error Rate metric (Snover et al., 2006) to determine the accuracy by measuring the number of edits to change a model output into the ground truth. We use their method to analyze the inference calibration.

5.2.1 Setup

We examine the impact of regularization effects induced by different target data on the model's output confidence and inference calibration for the main En→De tasks. We calculate the output confidence histograms and inference calibration errors for the En→De test set with the same settings of the multilingual models in Section 5.1.1. We plot the output confidence histograms (Figure 3) where the x-axis represents the output confidence scores and the y-axis represents the percentage of the number of tokens with those scores. In addition, we plot the reliability diagrams (Figure 4) to visualize the representations of model calibration where the x-axis is the average weighted confidence and the y-axis is the average weighted accuracy.

5.2.2 Discussion

First, **regularization from the small size of auxiliary target tasks improves inference calibration by penalizing output confidence.** For example, the main low-resource En→De translation task shows an over-confidence issue for its bilingual baseline model, see Figure 4a. The model seriously suffers from miscalibration, where the average gaps between confidence and accuracy are large (confidence $>$ accuracy). The small size of distant auxiliary target tasks can lead to better inference calibration. This regularization effect is achieved by penalizing over-confidence output (> 0.9) to enhance the model inference calibration, as shown in Figure 3a. These findings also align well with the medium-resource setting (1M). The relatively small



(a) InfECE=22.6 **(b)** InfECE=19.7 **(c)** InfECE=19.4 **(d)** InfECE=18.1 **(e)** InfECE=16.5

(f) InfECE=15.6 **(g)** InfECE=14.5 **(h)** InfECE=14.3 **(i)** InfECE=17.0 **(j)** InfECE=14.7

Figure 4: Reliability diagrams with inference calibration errors (InfECE) on the En→De test set in the low-resource (above) and medium-resource setting (below).

Main Task	Auxiliary Data	BLEU	△
	En→De	28.4	-0.2
En→De (4.5M)	En→Nl	28.3	-0.3
	En→Zh	29.0	+0.4

Table 3: The Main Task of En→De (4.5M) results with using Transformer-Big Model by adding 10% auxiliary tasks; △ represents the BLEU changes with the En→De baseline.

size of auxiliary target tasks (10%) benefits inference calibration from the penalizing over-confident output, shown in Figure 3b.

Second, **regularization from the large size of auxiliary target tasks improves inference calibration by improving translation accuracy.** Unlike in the small data (50%) scenario, which penalizes over-confident output probabilities to benefit the task, training with a large size of auxiliary target language pairs mainly helps the low-resource En→De task to improve translation accuracy to benefit inference calibration. Since similar language pairs share similar lexical and word order knowledge with the low-resource En→De task, they improve the accuracy more effectively.

5.3 Regularization in Larger Models

Section 5.1 and 5.2 show that utilizing small distant auxiliary data can benefit overfitting translation models from regularization, particularly for low- and medium-resource language pairs. For high-resource language pairs, Table 2 shows small distant data cannot help due to the "close-fitting" of the model parameters and training data. To further verify the impact of language regularization on high-resource language pairs, we increase the model size from Transformer-Base (93M) to Big (274M)[6] and utilize 10% of different auxiliary data to train with high-resource En→De translation task. Table 3 shows that 10% of distant auxiliary data En→Zh can help to improve around 0.6 BLEU points compared to the bilingual baseline while adding the same target languages or similar ones cannot. This finding further shows the effectiveness of language regularization for optimizing machine translation performance.

6 Conclusion

In this work, we disentangle the roles of knowledge transfer, source data size, and language regularization in one-to-many MT. In contrast with previous assumptions, we show that target-side knowledge transfer *does* play an important role in one-to-many MMT, which indicates that the increased amount of source data *is not* explain all the transfer. Future work can leverage this information to encourage different language pairs to have similar word representations to achieve the maximum positive transfer. Surprisingly, we find that using a small amount of linguistically distant auxiliary target data acts as an effective regularizer which results in translation performance gains. Such language regularization shows effectiveness in benefiting generalization ability and inference calibration. Our findings on language regularization shed new light on optimizing multilingual training by leveraging distant auxiliary data.

[6]For Transformer-Big, model details are shown in Appendix B, and the regular regularization techniques, e.g., dropout, we follow the same setup as (Vaswani et al., 2017).

7 Limitations

We acknowledge several limitations in our work. To directly understand the impact of knowledge transfer, source data, and regularization in one-to-many translation, we only observe the performance changes for one selected main language pair. Though translation results for auxiliary language pairs are provided in the Appendix D, further analysis of the dynamic performance trade-off between main and auxiliary language pairs is worthwhile to explore. Another limitation of our work is about the MMT setting, where we only work in one-to-many MT, while future work should extend it to many-to-many settings and explore the impact of adding multiple source languages.

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.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019.

Ting-Rui Chiang, Yi-Pei Chen, Yi-Ting Yeh, and Graham Neubig. 2021. Breaking down multilingual machine translation. In *Findings*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A comprehensive survey of multilingual neural machine translation. *ArXiv*, abs/2001.01115.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. 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 1723–1732, Beijing, China. Association for Computational Linguistics.

Patrick Fernandes, B. Ghorbani, Xavier García, Markus Freitag, and Orhan Firat. 2023. Scaling laws for multilingual neural machine translation. *ArXiv*, abs/2302.09650.

Luyu Gao, Xinyi Wang, and Graham Neubig. 2020. Improving target-side lexical transfer in multilingual neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3560–3566, Online. Association for Computational Linguistics.

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.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. 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 1565–1575, Hong Kong, China. Association for Computational Linguistics.

Surafel Melaku Lakew, Aliia Erofeeva, Matteo Negri, Marcello Federico, and Marco Turchi. 2018. Transfer learning in multilingual neural machine translation with dynamic vocabulary. In *International Workshop on Spoken Language Translation*.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, M. Zhang, and Tie-Yan Liu. 2021. R-drop: Regularized dropout for neural networks. *ArXiv*, abs/2106.14448.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. 2019. When does label smoothing help? In *Neural Information Processing Systems*.

Rudra Murthy, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2019. Addressing word-order divergence in multilingual neural machine translation for extremely low resource languages. 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 3868–3873, Minneapolis, Minnesota. Association for Computational Linguistics.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. 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.

Uri Shaham, Maha Elbayad, Vedanuj Goswami, Omer Levy, and Shruti Bhosale. 2022. Causes and cures for interference in multilingual translation.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

David Stap, Vlad Niculae, and Christof Monz. 2023. Viewing knowledge transfer in multilingual machine translation through a representational lens. *ArXiv*, abs/2305.11550.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

Y. Tang, C. Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *ArXiv*, abs/2008.00401.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020. On the inference calibration of neural machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Yining Wang, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2018. Three strategies to improve one-to-many multilingual translation. In *Conference on Empirical Methods in Natural Language Processing*.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. 2019. Characterizing and avoiding negative transfer. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11285–11294.

Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big Data*, 3(1):9.

Yilin Yang, Akiko Eriguchi, Alexandre Muzio, Prasad Tadepalli, Stefan Lee, and Hany Hassan. 2021. Improving multilingual translation by representation and gradient regularization. *ArXiv*, abs/2109.04778.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

A Language Choices

Table 4 shows the linguistic information about the main and auxiliary target languages. Table 5 shows two linguistic groups trained with the main language pair.

ISO	Lang.	Family	Script
De	**German**	**Germanic**	**Latin**
Nl	Dutch	Germanic	Latin
Et	Estonia	Uralic	Latin
Ru	Russian	Slavic	Cyrillic
Zh	Mandarin	Chinese	Chinese
Es	**Spanish**	**Romance**	**Latin**
Pt	Portuguese	Romance	Latin
Nl	Dutch	Germanic	Latin
Ru	Russian	Slavic	Cyrillic
Zh	Mandarin	Chinese	Chinese
Ru	**Russian**	**Slavic**	**Cyrillic**
Uk	Ukrainian	Slavic	Cyrillic
Cs	Czech	Slavic	Latin
De	German	Germanic	Latin
Zh	Mandarin	Chinese	Chinese

Table 4: The linguistic information for the main and auxiliary target languages. **Bold** designates the main target languages: De, Es, Ru.

ISO	Lang.	Family	Script
Af	Afrikaans	Germanic	Latin
Da	Danish	Germanic	Latin
Nl	Dutch	Germanic	Latin
Is	Icelandic	Germanic	Latin
No	Norwegian	Germanic	Latin
Sv	Swedish	Germanic	Latin
Gl	Galician	Romance	Latin
Es	Spanish	Romance	Latin

ISO	Lang.	Family	Script
Bg	Bulgarian	Slavic	Cyrillic
Cs	Czech	Slavic	Cyrillic
Mk	Macedonian	Slavic	Cyrillic
Pl	Polish	Slavic	Cyrillic
Sr	Serbian	Slavic	Cyrillic
Sk	Slovak	Slavic	Cyrillic
Sl	Slovenian	Slavic	Cyrillic
Uk	Ukrainia	Slavic	Cyrillic

Table 5: Two groups of auxiliary target languages.

B Model Parameters

We follow the setup of the Transformer-base and Transformer-big models (Vaswani et al., 2017). For each model, the number of layers in the encoder and in the decoder is $N = 6$. For Transformer-base, we employ $h = 8$ parallel attention layers or heads. The dimensionality of input and output is $d_model = 512$, and the inner layer of feed-forward networks has dimensionality $d_{ff} = 2048$. For Transformer-big, we employ $h = 16$ parallel attention layers or heads. The dimensionality of input and output is $d_model = 1024$, and the inner layer of feed-forward networks has dimensionality $d_{ff} = 4096$.

C Dataset Statistics

The data statistics of main language pairs are shown in Table 6. The data statistics of joint training target language pairs are shown in Table 7.

Language	ISO	Dataset Source	Validation Set	Test Set
German	De	WMT14	WMT14	WMT14
Spanish	Es	WMT13	WMT13	WMT13
Russian	Ru	WMT22	WMT22	WMT22

Table 6: The data statistics of main low- and medium-resource language pairs. For each language, we display the ISO code, language name, sampled training dataset source, validation set, and test set. Sampled training low-resource dataset size: 100K, sampled training medium-resource dataset size: 1M.

Language	ISO	Dataset Source	Validation/Test Set
Estonia	Et	WMT18	WMT18
Chinese	Zh	WMT19	WMT19
Portuguese	Pt	WMT16	WMT16
Ukrainian	Uk	WMT22	WMT22
Czech	Cs	WMT22	WMT22
Dutch	Nl	CCMatrix	CCMatrix
Afrikaans	Af	CCMatrix	CCMatrix
Danish	Da	CCMatrix	CCMatrix
Icelandic	Is	CCMatrix	CCMatrix
Norwegian	No	CCMatrix	CCMatrix
Swedish	Sw	CCMatrix	CCMatrix
Galician	Gl	CCMatrix	CCMatrix
Bulgarian	Bg	CCMatrix	CCMatrix
Macedonian	Mk	CCMatrix	CCMatrix
Polish	Pl	CCMatrix	CCMatrix
Serbian	Sr	CCMatrix	CCMatrix
Slovak	Sk	CCMatrix	CCMatrix
Slovenian	Sl	CCMatrix	CCMatrix

Table 7: The data statistics of auxiliary training target language pairs. For each language, we display the ISO code, language name, sampled training dataset source, and validation set. The validation and test sets from CCMatrix, are randomly sampled from the CCMatrix corpus, each containing 2000 samples.

D Additional Results

Here, we show all auxiliary language BLEU scores in Table 8, 9 and 10.

		En→De		
α%	en→nl	en→et	en→ru	en→zh
10%	$8.9_{0.2}$	$6.2_{0.7}$	$6.0_{0.6}$	$5.5_{0.5}$
50%	$11.9_{0.2}$	$11.2_{0.3}$	$10.2_{0.3}$	$9.8_{0.3}$
100%	$20.3_{0.2}$	$11.9_{0.4}$	$13.7_{0.2}$	$12.3_{0.4}$
500%	$23.7_{0.3}$	$14.3_{0.1}$	$17.6_{0.3}$	$15.6_{0.2}$
1000%	$26.4_{0.2}$	$15.3_{0.5}$	$18.5_{0.1}$	$16.7_{0.3}$
		En→Ru		
α%	en→uk	en→cs	en→de	en→zh
10%	$8.8_{0.6}$	$7.6_{0.6}$	$7.8_{0.2}$	$5.0_{0.2}$
50%	$15.0_{0.5}$	$12.2_{0.2}$	$10.2_{0.3}$	$9.3_{0.1}$
100%	$18.3_{0.3}$	$12.6_{0.1}$	$11.0_{0.2}$	$12.5_{0.4}$
500%	$22.7_{0.2}$	$14.2_{0.1}$	$16.8_{0.2}$	$15.1_{0.1}$
1000%	$23.4_{0.1}$	$14.7_{0.2}$	$18.9_{0.2}$	$16.2_{0.2}$
		En→Es		
α%	en→pt	en→nl	en→ru	en→zh
10%	$9.2_{0.4}$	$8.6_{0.6}$	$6.2_{0.3}$	$5.1_{0.8}$
50%	$12.3_{0.3}$	$11.3_{0.6}$	$10.0_{0.2}$	$9.2_{0.3}$
100%	$20.5_{0.3}$	$15.2_{0.3}$	$11.5_{0.3}$	$12.5_{0.2}$
500%	$23.2_{0.2}$	$18.2_{0.3}$	$16.5_{0.3}$	$15.6_{0.2}$
1000%	$26.2_{0.4}$	$19.6_{0.1}$	$18.6_{0.3}$	$16.4_{0.1}$

Table 8: BLEU scores for the auxiliary language pairs in a low-resource setting (100K) when training with main language pairs: En→De, En→Es, and En→Ru. $\alpha\% = 10, 50, 100, 500, 1000$ represents the proportion of the low-resource setting size.

		En→De		
α%	en→nl	en→et	en→ru	en→zh
1%	$12.6_{0.2}$	$7.0_{0.7}$	$7.0_{0.6}$	$6.7_{0.5}$
10%	$22.7_{0.2}$	$12.3_{0.3}$	$12.7_{0.3}$	$13.5_{0.3}$
50%	$25.5_{0.2}$	$16.0_{0.4}$	$17.8_{0.2}$	$16.7_{0.4}$
100%	$28.4_{0.3}$	$16.5_{0.1}$	$18.2_{0.3}$	$16.5_{0.2}$
200%	$29.4_{0.0}$	$15.0_{0.0}$	$18.1_{0.0}$	$16.4_{0.0}$
		En→Ru		
α%	en→uk	en→cs	en→de	en→zh
1%	$13.8_{0.6}$	$8.2_{0.6}$	$7.0_{0.2}$	$5.8_{0.2}$
10%	$18.0_{0.5}$	$11.2_{0.2}$	$12.5_{0.3}$	$12.3_{0.1}$
50%	$20.3_{0.3}$	$12.6_{0.1}$	$16.0_{0.2}$	$16.5_{0.4}$
100%	$23.7_{0.2}$	$15.2_{0.1}$	$17.8_{0.2}$	$16.1_{0.1}$
200%	$26.4_{0.0}$	$16.7_{0.0}$	$19.9_{0.0}$	$16.2_{0.0}$
		En→Es		
α%	en→pt	en→nl	en→ru	en→zh
1%	$12.2_{0.4}$	$10.6_{0.6}$	$7.2_{0.3}$	$6.1_{0.8}$
10%	$19.3_{0.3}$	$12.3_{0.6}$	$13.0_{0.2}$	$14.2_{0.3}$
50%	$22.5_{0.3}$	$19.2_{0.3}$	$17.5_{0.3}$	$16.5_{0.2}$
100%	$27.2_{0.2}$	$20.2_{0.3}$	$18.5_{0.3}$	$16.6_{0.2}$
200%	$28.2_{0.0}$	$20.2_{0.0}$	$18.6_{0.0}$	$16.0_{0.0}$

Table 9: BLEU scores for the auxiliary language pairs in a mid-resource setting (1M) when training with main language pairs: En→De, En→Es, and En→Ru. $\alpha\% = 1, 10, 50, 100, 200$ represents the proportion of the medium-resource setting size.

		En→De		
α%	en→nl	en→et	en→ru	en→zh
1%	14.0	9.3	7.6	8.9
10%	24.1	14.5	15.8	16.5
50%	24.4	17.0	16.2	17.0
100%	25.0	19.5	15.7	16.5
200%	25.6	20.1	14.1	15.0

Table 10: BLEU scores for the auxiliary language pairs in a high-resource setting (4.5M) when training with main language pairs: En→De. $\alpha\% = 1, 10, 50, 100, 200$ represents the proportion of the high-resource setting size.