Language Tokens: A Frustratingly Simple Approach Improves Zero-Shot Performance of Multilingual Translation

Muhammad ElNokrashy muelnokr@microsoft.com
Amr Hendy amrhendy@microsoft.com
Mohamed Maher mohamedmaher@microsoft.com
Mohamed Afify mafify@microsoft.com
Microsoft ATL, Cairo

Hany Hassan Awadalla hanyh@microsoft.com
Microsoft, Redmond

Abstract

This paper proposes a simple yet effective method to improve direct (*X-to-Y*) translation for both cases: zero-shot and when direct data is available. We modify the input tokens at both the encoder and decoder to include signals for the source and target languages. We show a performance gain when training from scratch, or finetuning a pretrained model with the proposed setup. In the experiments, our method shows nearly 10.0 BLEU points gain on in-house datasets depending on the checkpoint selection criteria. In a WMT evaluation campaign, *From-English* performance improves by 4.17 and 2.87 BLEU points, in the zero-shot setting, and when direct data is available for training, respectively. While *X-to-Y* improves by 1.29 BLEU over the zero-shot baseline, and 0.44 over the many-to-many baseline. In the low-resource setting, we see a 1.5 $\sim$1.7 point improvement when finetuning on *X-to-Y* domain data.

1 Introduction

Neural machine translation (**NMT**) has witnessed significant advances since the introduction of the transformer model (Vaswani et al., 2017). This model has shown impressive performance for bilingual translation commonly from and to English (Hassan et al., 2018). It has also been shown that the proposed model could be easily extended to multiple language pairs (Aharoni, Johnson, & Firat, 2019; Fan et al., 2020; Johnson et al., 2017; X. Wang, Tsvetkov, & Neubig, 2020), to and/or from English, by simple modifications to the basic architecture. This holds promise for improved performance for low-resource pairs through transfer learning, as well as better training and deployment costs per language pair. This setting is referred to as multilingual neural machine translation (**MNMT**).

The mainstream method of training MNMT is to introduce an additional input tag at the encoder to indicate the target language, while the decoder uses the usual begin-of-sentence (`BOS`) token. This simple modification to the bilingual architecture is shown to work well up to hundreds of language pairs (Fan et al., 2020; Tran et al., 2021), given a corresponding increase in the number of parameters to handle the increased training data. Despite the emergence of modified architectures which add language-specific parameters, like language specific subnetworks (LASS) (Lin, Wu, Wang, & Li, 2021), and adapters (Bapna & Firat, 2019), the basic architecture remains the most effective choice for deploying large scale production systems.

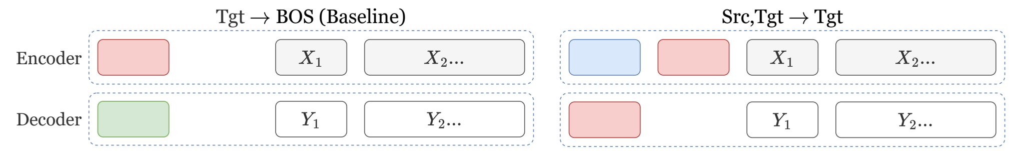


Figure 1: Comparing tokens as seen by the Encoder and the Decoder in the *(Left)* baseline (`T-B`) and in the *(Right)* top proposed method (`ST-T`).

2 Motivation

While MNMT was originally focused on English-centric translation, there is increasing interest interest in direct translation[1] rather than pivoting through a common language (ex. English). In Freitag and Firat (2020), the authors mine direct translation data by matching the English part of English-centric corpora, then use modified temperature sampling to alleviate the overrepresentation of English as target. Another work (Fan et al., 2020) leverages public direct translation data including CCMatrix (Schwenk, Wenzek, Edunov, Grave, & Joulin, 2019) and CCAligned (El-Kishky, Chaudhary, Guzman, & Koehn, 2019) as well as improved sampling and sparse modeling to develop a 100-by-100 direct translation model.

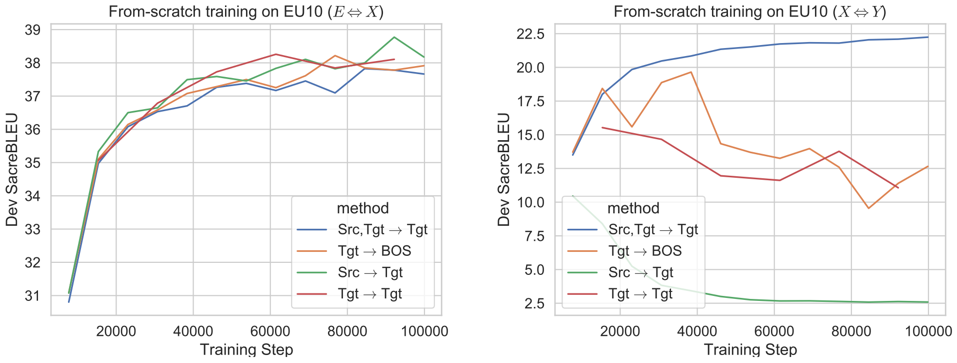
While the availability of direct translation training data helps improve the corresponding directions, the zero-shot case remains of particular importance considering the difficulty in maintaining coverage of the set of rapidly increasing directions, and handling the corresponding increase in data size and compute time and resources. Hence the interest in techniques that improve zero-shot translation performance, benefit from parallel training data as it becomes available, and which can be easily applied to the basic architecture and pretrained models.

Related Works. Yang et al. (2021) approaches the off-target translation problem using gradient projection and no direct training data. Zhang, Williams, Titov, and Sennrich (2020) improves the zero-shot case using online back-translation and by specializing layers (ex. LayerNorm) for the target language. Rios, Müller, and Sennrich (2020) utilizes separately-trained vocabularies per language. Arivazhagan et al. (2019) proposes an alignment loss to enforce source language invariance. Ha, Niehues, and Waibel (2016) proposes tagging input tokens by the source language and indicating the target language directly to a shared decoder.

Proposal. We propose a simple yet effective method that improves the performance of direct translation: The input tokens used in MNMT are changed to `ST-T` instead of `T-B` (see Figure 1). The encoder takes tokens for both the source and target languages (`S,T`) while the decoder takes one for only the target language (`T`). It is shown that using these modified tokens significantly improves the performance on direct translation pairs without any parallel $X \Leftrightarrow Y$ translation data—training only on English-centric data ($E \Leftrightarrow X$). Remarkably, these gains are quickly obtained if we start from a model trained using the baseline tokens and continue training after adding the new tokens. In subsequent experiments, we also show that some gains are still observed if we continue training the baseline model using a mix of direct ($X \Leftrightarrow Y$) and English-centric training data—suggesting the method extends to the non zero-shot case as well.

The paper is organized as follows: We describe the proposed method in Section 3. This is followed by describing the data and the models used in our experiments in Section 4 and Section 5 respectively. Section 6 gives the experimental results (domain finetuning results in Section 6.2). Finally, Section 7 gives the conclusion.

[1] Also known as *X-Y* translation. In the rest of the paper we refer to translation between any two language pairs not involving English as *direct* or *X-Y* translation.



(a) SacreBLEU for $E \Leftrightarrow X$ dev.

(b) SacreBLEU for $X \Leftrightarrow Y$ dev.

Figure 2: **Validation for from-scratch training.** All training uses English-centric data and no direct ($X \Leftrightarrow Y$) data. The baseline (`T-B`) quickly loses performance on $X \Leftrightarrow Y$ dev. The `S-T` method fails on $X \Leftrightarrow Y$ dev, but matches or exceeds alternatives on $E \Leftrightarrow X$ dev. The proposed method (`ST-T`) matches the alternatives on $E \Leftrightarrow X$ dev (within 1.0 BLEU) and maintains high performance for $X \Leftrightarrow Y$ generalization.

3 Approach

In basic MNMT, the source sentence is followed by a token indicating the target language at the encoder side, and with the begin-of-sentence token at the decoder side. This setup is `T-B` in **Figure 1**. In the proposed method, we perform a simple modification by adding both the source and target tokens to the input at the encoder side[2], and the target at the decoder side. This setup is `ST-T` in Figure 1. Y. Wang, Zhang, Zhai, Xu, and Zong (2018) shows that adding the `TGT` language to the decoder input helps English-to-X translation. In a recent submission to WMT21, Tran et al. (2021) uses a `SRC` token at the encoder and a `TGT` token at the decoder, which can be observed from the public evaluation code[3]. In initial experiments we try several variants of indicating the languages to the model. We find that most are similar for the English-centric case, but the proposed method (`ST-T`) performs the best for zero-shot direct translation.

3.1 Initial Experiments

To validate the proposed method, we train a model on 10 European languages using in-house English-centric data—Once using the baseline tokens, and once using the new tokens. Details of the model and the training are given in Section 5. The graph of the dev BLEU score during training is shown for the English-centric devset and the direct devset in **Figures 2a** and **2b**. Also shown in the figure is the `S-T` setup similar to (Tran et al., 2021), and the `T-T` setup which passes the target language to both encoder and decoder.

3.2 Language Coding and Model Conditioning

While all models perform similarly for the English-centric set on which they are trained, the behavior is different for the novel direct set. The baseline and proposed models are close at the beginning of the training but the baseline quickly deteriorates *as it improves in its assigned task* on English-centric data. One explanation is that the conditioning in the `T-B` case explicitly

[2] Note that the order of source and target is not significant and that we also add the target at the decoder.

[3] Found at: https://github.com/facebookresearch/fairseq/blob 47c58f0858b5484a18f39549845790267cffee1a/examples/wmt21/eval.sh

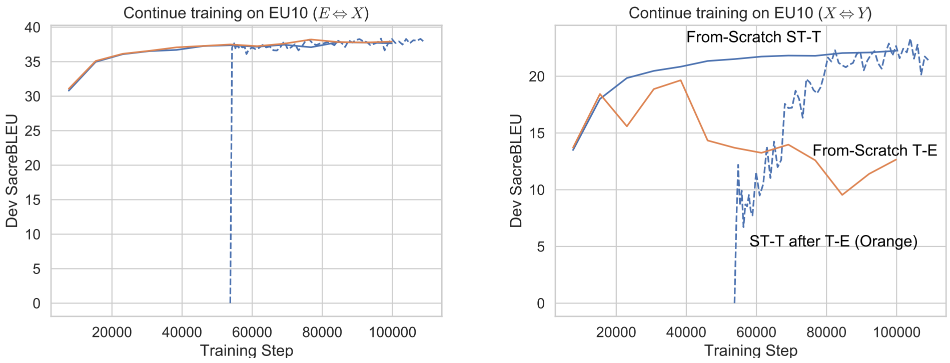
Directions	`T-B`	`ST-T`
$E \Leftrightarrow X$	0.14%	0.15%
$X \Leftrightarrow Y$	29.36%	1.69%
Both	23.49%	**1.35**%

Table 1: Percentage of *off-target* samples in the $E \Leftrightarrow X$ and $X \Leftrightarrow Y$ dev sets for EU10 measured at 53k steps. Training graphs in Figure 2.

indicates only the target language, while the source language is inferred. Consider also the absence of direct $X \Leftrightarrow Y$ training data, then an *implicit, and valid, pattern* emerges: When `TGT≠en`, it is implied that `SRC=en`. Thus, in the $X \Leftrightarrow Y$ test case, the model expects the source to be in English, which may be a source of confusion. Conversely, the model with the `S-T` setup performs very poorly from the beginning for the direct set. We propose that this is in line with findings that show that encoder capacity may be of higher importance to MT than decoder capacity (Kasai, Pappas, Peng, Cross, & Smith, 2020; Kim et al., 2019). Removing the `TGT` signal from the encoder would then be a significant handicap. It may work in the English-centric case because the *implicit pattern* described above is sufficient conditioning.

Off-Target Translations. **Table 1** shows Language ID mis-matches for `T-B` and `ST-T` using fasttext language identification on the *from-scratch* experiments on EU10 (fig. 2).

3.3 Building on Pretrained Models



(a) SacreBLEU for $E \Leftrightarrow X$ dev.

(b) SacreBLEU for $X \Leftrightarrow Y$ dev.

Figure 3: Orange is the baseline (`T-B`) setup, while blue is the proposed (`ST-T`) setup. The solid lines start from scratch. The dashed line continues training from step 53k of the baseline, using `ST-T` tokens, on English-centric ($E \Leftrightarrow X$) data. No direct $X \Leftrightarrow Y$ data is used for training. $E \Leftrightarrow X$ performance is quickly regained. $X \Leftrightarrow Y$ performance approaches that of from-scratch training within a similar *total* training budget (dashed vs. solid blue lines).

An interesting scenario is how to make use of already trained models that would be costly to retrain. We validate our top proposed method in that case by continuing training from a midway checkpoint of the baseline. The performance on both dev sets is shown in **Figure 3**. Performance (dashed blue line) on the $E \Leftrightarrow X$ set starts at zero but rapidly regains its baseline value, then remains steady as training progresses. The same happens for $X \Leftrightarrow Y$ data but at a slower pace. The weights already trained on English-centric data seem to realign to the new setup efficiently.

To summarize: Compared to the baseline, the proposed tokens perform similarly on English-centric tests and significantly outperform on direct translation tests while training using English-centric data in both cases. If we continue training the baseline model by adding the proposed tokens we quickly recover the performance of both the English-centric and direct data to levels obtained by training from scratch, as seen in the initial experiments. Therefore, in the rest of this paper we focus on continuing training a model pretrained using the baseline tokens.

4 Data

In initial experiments we build a model for 10 European languages using in-house data (Section 4.1). Follow-up experiments use WMT data (Akhbardeh et al., 2021) as well as other publicly available data covering 6 languages (Section 4.2). Most experiments use English-centric training data, some use direct training data (Section 4.3), and some use domain data (Section 4.5). Validation data is described in Section 4.4. Tables are in Appendix A.

4.1 EU10 Training Data

EU10 is an in-house web-crawled parallel dataset with a total of 3.35 billion sentence pairs covering 10 European languages: Dutch (nl), English (en), French (fr), German (de), Greek (el), Italian (it), Polish (pl), Portuguese (pt), Spanish (es), and Romanian (ro). Details in Table 8.

4.2 WMT Training Data

For WMT, we use the data for 12 English-centric language pairs provided by the news translation shared task in WMT21[4], and additionally data from the public sources CCMatrix (Schwenk et al., 2019) and CCAligned (El-Kishky et al., 2019). The combined set covers the directions of English (en) to and from: Czech (cs), German (de), Icelandic (is), Japanese (ja), Russian (ru), and Chinese (zh). We apply some preprocessing steps to filter noisy data. We filter for the expected languages using fasttext (Joulin, Grave, Bojanowski, & Mikolov, 2017); normalize punctuation using moses[5]; then discard sentences longer than 250 words or with a source/target or target/source length ratio exceeding 3. The filtered data totals 2.16 billion sentence pairs. Details on filtered parallel data sizes are shown in Table 10. Note the difference in the parallel data of English-centric directions for the same non-English language is due to having different amounts of synthetic data released by the WMT21 shared task.

4.3 Direct Training Data

For experiments with direct $X \Leftrightarrow Y$ data covering the 7 languages in the WMT dataset (Section 4.2), we build a dataset of parallel training data in 42 translation directions, including the English-centric directions. We collect $X \Leftrightarrow Y$ data from publicly available sources as described in Section 4.2. The resulting size of the collected bitext $X \Leftrightarrow Y$ data is shown in Table 9. We then sample 10 million sentence pairs from the WMT English-centric dataset for each direction to avoid catastrophic forgetting on $E \Leftrightarrow X$ directions. We end up with a many-to-many dataset with a total of 525 millions sentence pairs that contains 405 millions sentence pairs in $X \Leftrightarrow Y$ directions, and 120 millions sentence pairs in English-centric directions.

4.4 Development and Test Data

We evaluate the translation performance on various devsets depending on the training dataset and the language list. For experiments using in-house training data (Section 4.1), we use in-

[4] https://www.statmt.org/wmt21/translation-task.html (Akhbardeh et al., 2021).
[5] https://github.com/moses-smt/mosesdecoder (Koehn et al., 2007).

house dev sets covering all the many-to-many 90 translation directions. For experiments using the WMT data described in Section 4.2 and 4.3, we use publicly available dev sets composed of: WMT21-provided sets to cover the 12 English-centric directions, and the FLORES-101 benchmark (Goyal et al., 2022) to cover the remaining 30 $X \Leftrightarrow Y$ directions.

4.5 Data for Domain Experiments

We utilize domain data from the OPUS project[6]. We collect the EMEA, JRC-ACQUIS and TANZIL domains in the directions German to/from Czech. The data is shuffled and split into train, test and validation sets. Any sentences that occur in the validation or test sets are removed from training. The sizes of the splits are shown in Table 7. EMEA is a parallel corpus of PDF documents from the European Medicines Agency. JRC-ACQUIS is a collection of legislative text of the European Union that comprises selected texts written between the 1950s and now. TANZIL is a collection of Quran translations compiled by the Tanzil project.

5 Models

All experiments use the same architecture and configuration. We use the Transformer encoder-decoder architecture (Vaswani et al., 2017) as the base model and opt for a deep encoder and a shallower decoder as presented in Kim et al. (2019) and Kasai et al. (2020), with 24 encoder layers and 12 decoder layers. Dimensions are 1024 for model width, 4096 for the feed-forward hidden layer, and 16 attention heads. We use pre-layer normalization which is becoming more common for similar architectures (Xiong et al., 2020). We use a vocabulary of size $128,000$ with the sentencepiece tokenizer[7]. The model size is 0.6B parameters. All models are trained by the RAdam optimizer (Liu et al., 2019). See Appendix A, Table 6 for other hyper-parameters.

6 Experimental Results

In this section, we show experimental results using the WMT model as described in Section 5. We first continue training the WMT model using the proposed tokens on English-centric data only, then we continue training using direct data[8]. In other experiments, we use $X \Leftrightarrow Y$ data from start, once with each of `T-B` and `ST-T` tokens. In all cases, we report the BLEU score for both the direct and English-centric dev sets. See Table 2. The results of continuing training the baseline tokens with direct data are shown in the fourth row (`D` *Direct FT*). Note that the third row (`P-D` *Proposed* $\hookrightarrow$ *Direct FT*) corresponds to continuing training with new tokens for 47k iterations, then adding the direct data—running for 112k steps in total.

6.1 Medium-resource MNMT

In **Table 2**, the first row shows the scores of the base WMT model on both English-centric and direct dev sets (in zero-shot setting) with the `T-B` setup. Continuing to train the model using the new tokens shows gains on both dev sets, although less than what would be expected from the initial results on EU10 (Section 3.1). Continuing to train the base model using direct data shows larger gains on the direct dev set, but a smaller gain on English-centric dev.

The third and last rows show that we still observe some gain from the new tokens after adding the direct data. The best strategy (row 3) is two phases: to train first using the English-centric data, then add the direct data. Consider Figure 4b: It may require a smaller $X \Leftrightarrow Y$ dataset (fewer steps) in the two-phase setup than when using direct data from the start.

[6] https://opus.nlpl.eu/index.php (Tiedemann, 2012).

[7] https://github.com/google/sentencepiece (Kudo & Richardson, 2018).

[8] This is a mix of $X \Leftrightarrow Y$ and $E \Leftrightarrow X$ data (Section 4.3) to avoid catastrophic forgetting of English-centric translation.

[9] SacreBLEU signature: `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.1.0`. For targets in Chinese and Japanese, the tokenizers used are `zh` and `ja-mecab`. (Post, 2018)

#	Setup	**Direct**	English-Centric	$X \Rightarrow E$	$E \Rightarrow X$	Best At
B	Base Model	13.67	26.30	27.37	25.28	-
P	↪ Proposed	14.96	**30.27**	**31.10**	**29.45**	63k
P-D	↪ Direct FT (from 47k)	**23.59**	*28.83*	*29.52*	*28.15*	112k
D	↪ Direct FT	23.15	28.10	28.90	27.30	120k
DP	↪ Proposed & Direct FT	23.09	28.19	28.85	27.52	102k

Table 2: **SacreBLEU**[9] of the base WMT model (first row) when finetuned in various setups. A hooked arrow (↪) indicates a row that continues training from the parent model. Thus rows `P`, `D`, and `DP` show finetuning using the proposed (`ST-T`) tokens, direct data, or both; while row `P-D` continues from row `P`. *Direct FT* refers to finetuning with direct data. The ***Direct*** column uses FLORES-100 dev, while the *English-centric* column uses WMT21 dev. The *Best At* column reports the *total* training steps starting from *Base Model*.

#	Setup	**Direct**	English-Centric	$X \Rightarrow E$	$E \Rightarrow X$	Best At
B	Base Model	-16.23	36.00	32.69	39.31	-
P	↪ Proposed	6.36	**49.87**	**45.21**	**54.54**	63k
P-D	↪ Direct FT (from 47k)	**54.30**	*47.14*	*44.18*	*50.10*	112k
D	↪ Direct FT	51.74	44.71	41.73	47.69	120k
DP	↪ Proposed & Direct FT	51.57	44.47	41.82	47.12	102k

Table 3: Average **COMET**[10] (×100) of the set of WMT experiments. Same notation as Table 2.

Table 3 shows the COMET scores for the same experiments set. While both metrics agree in rankings, COMET suggests larger gains than suggested by BLEU.

See **Table 4** for COMET-based statistical significance model comparison aggregates across language directions.

See **Figure 4** for training performance curves for rows 1–4. Consider that row `P-D` sees only $E \Leftrightarrow X$ for the first phase (corresponding to `P`), and then sees a small amount of $X \Leftrightarrow Y$ data before improving. It may thus help in making better use of a smaller $X \Leftrightarrow Y$ dataset.

6.2 Low-resource Adaptation

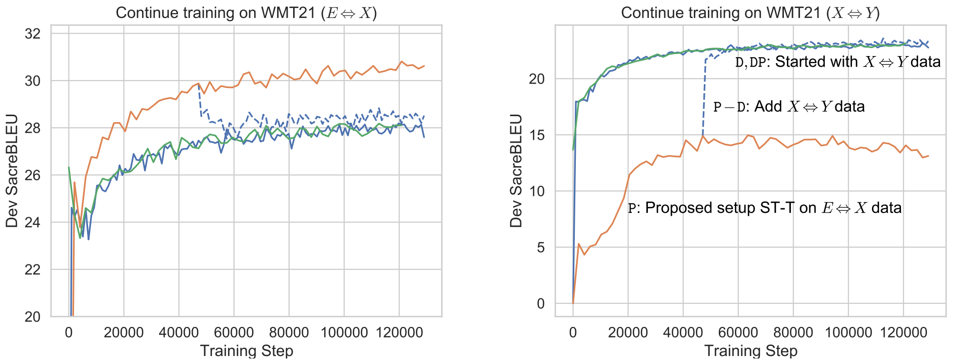
For the low-resource setting, we use a domain adaptation example. We start from the WMT model with the baseline and new tokens (corresponding to the first and second rows of Table 2). We finetune a separate model for each adaptation experiment: for German to and from Czech, and for the domains EMEA, JRC and Tanzil as obtained from OPUS. Details of the data are found in Table 7. The results for CS-DE and DE-CS are shown in Table 5.

From **Table 5**, the new tokens improve both the pretrained and the finetuned models. The difference depends on the direction and the domain but is generally noticeable. This is an interesting scenario because we can start from an English-centric baseline and continue training using the new tokens to create a stronger base model that improves downstream performance for different directions and domains.

[10] COMET model `wmt20-comet-da` version `1.1.1` (Rei, Stewart, Farinha, & Lavie, 2020).

Direction	Without $X \Leftrightarrow Y$ train data		With $X \Leftrightarrow Y$ train data	
	`B` Base Model	`P` Proposed	`D` Direct FT	`P-D` Proposed ↪ Direct
$X \Rightarrow E$	0	**6**	0	**4**
$E \Rightarrow X$	0	**6**	0	**3**
$X \Leftrightarrow Y$	10	**20**	0	**10**
Total/42	10	**32**	0	**17**

Table 4: To compare models to the nearest baseline, we calculate statistical significance with the Paired T-Test and bootstrap resampling at $p < 0.05$, following Koehn (2004). Each cell shows the count of *wins* for a model and direction. In the zero-shot setting, the proposed method outperforms the baseline in $32/42$ directions. With direct parallel data available, the proposed method outperforms the continued training baseline in $17/42$ directions, and is tied in the rest.



(a) SacreBLEU for $E \Leftrightarrow X$ dev.

(b) SacreBLEU for $X \Leftrightarrow Y$ dev.

Figure 4: All solid line models start from the **pretrained base model** (row `B` in Tables 2 & 3). Orange is the proposed setup trained on $E \Leftrightarrow X$ data (row `P`). Blue adds $X \Leftrightarrow Y$ data (rows `P-D` & `DP`) where: Run `P-D` (dashed line) continues training on $X \Leftrightarrow Y$ data from step 47k of run `P`, while run `DP` (solid blue) starts with that data. Note that `D` and `DP` perform similarly on both dev sets, but `P-D` improves $X \Leftrightarrow Y$ performance while lessening the loss of $E \Leftrightarrow X$ performance that is gained from run `P`. `P-D` reaches similar performance to `D` and `DP` in fewer steps with access to $X \Leftrightarrow Y$ data directly, suggesting improved data efficiency.

7 Conclusion

This paper proposes a simple and effective method to improve direct translation for both the zero-shot case and when direct data is available. The input tokens used for MNMT are changed from `T-B` (encoder → decoder) to `ST-T`. Moreover, the performance of the new tokens can be readily obtained if we continue training the baseline model with the new tokens but the same training data. For a WMT-based setting, we see around 1.3 BLEU points improvement for zero-shot direct translation and around 0.4 BLEU point improvement when using direct data for training. In both cases the English-centric performance is also improved—by as much as 3.97 in WMT21 for one setup. COMET scores see noticeable bumps as well—by 2.56 points. on $X \Leftrightarrow Y$ dev, and 15.23 points on $E \Rightarrow X$ dev. On another front, the proposed tokens are effective when finetuning a general model for direct translation using domain data.

For three tested domains and two translation directions, we see significant improvements over the baseline. Results for EU10 (Section 3.1) suggest a stronger potential given more similar language and domain sets.

Model	Domain	CS-DE			DE-CS		
		`B` Base	`P` Proposed	*Delta*	`B` Base	`P` Proposed	*Delta*
Zero-Shot	EMEA	35.2	35.3	+0.1	36.9	39.5	+2.6
	JRC	45.0	48.0	+3.0	45.1	47.6	+2.5
	Tanzil	6.6	10.5	+3.9	6.5	9.7	+3.2
Finetuned	EMEA	45.8	46.4	+0.6	46.2	48.2	+2.0
	JRC	53.7	56.0	+2.3	52.7	54.5	+1.8
	Tanzil	24.4	26.0	+1.6	26.0	27.2	+1.2
Zero-Shot	*Average*	28.9	31.3	+**2.4**	29.5	32.3	+**2.8**
Finetuned		41.3	**42.8**	+1.5	41.6	**43.3**	+1.7

Table 5: Results of finetuning on different domains using the baseline and proposed tokens for Czech from and to German. The model is finetuned separately for each domain and direction.

A Appendix

Parameter	WMT			EU10
	Pretraining	Finetuning	Domain adaptation	
Optimizer	RAdam	RAdam	RAdam	RAdam
Learning Rate	0.001	0.008	0.00089	0.015
LR Scheduler	Inverse Sqrt	Inverse Sqrt	Inverse Sqrt	Inverse Sqrt
Warmup	$4,000$	$5,000$	800	$5,000$
Batch Size	0.8M	1.5M	1M	2M

Table 6: Hyper-parameters comparison between experiment sets. The LR values were not optimized for these experiments, but inherited from unrelated trials. Note that between any two *phases* of an experiment (for example in `P-D`, adding $X \Leftrightarrow Y$ data in the second phase), all non-parameter state is re-initialized, including LR scheduler and optimizer state.

Domain	Training Set Size	Validation Set Size	Test Set Size
EMEA	1.06M	561	582
JRC-acquis	1.15M	609	1,190
Tanzil	45k	326	302

Table 7: Sentence counts of train, development, and test sets for domain data.

Language pair (XE)	**# sentences** (M)	**Language pair** (EX)	**# sentences** (M)
Dutch → English	195	English → Dutch	233
French → English	298	English → French	251
German → English	250	English → German	219
Greek → English	166	English → Greek	117
Italian → English	237	English → Italian	170
Polish → English	175	English → Polish	161
Portuguese → English	108	English → Portuguese	64
Spanish → English	260	English → Spanish	171
Romanian → English	162	English → Romanian	112

Table 8: In-house web crawled parallel data statistics used in EU10 training. We report the list of 18 language directions and the number of sentences (Millions) per each language pair.

Language pair (XY)	**# sentences** (M)	**Language pair** (XY)	**# sentences** (M)
Czech ↔ German	33	German ↔ Chinese	19
Czech ↔ Icelandic	0.6	Icelandic ↔ Japanese	1.1
Czech ↔ Japanese	11	Icelandic ↔ Russian	2.1
Czech ↔ Russian	28	Icelandic ↔ Chinese	0.7
Czech ↔ Chinese	6.6	Japanese ↔ Russian	9.5
German ↔ Icelandic	3.4	Japanese ↔ Chinese	12.4
German ↔ Japanese	15	Russian ↔ Chinese	14
German ↔ Russian	46		

Table 9: Bitext data for 30 X→Y language directions collected from CCMatrix and CCAligned. We report the number of sentences (Millions) per each language pair.

Language pair (XE)	**# sentences** (M)		**Language pair** (EX)	**# sentences** (M)	
	Raw	**Cleaned**		**Raw**	**Cleaned**
Czech → English	206	189	English → Czech	181	165
German → English	436	411	English → German	436	411
Icelandic → English	15	13.4	English → Icelandic	15	13.4
Japanese → English	85	81	English → Japanese	85	81
Russian → English	289	273	English → Russian	292	280
Chinese → English	139	132	English → Chinese	119	113

Table 10: Bitext data includes data released by the WMT21 shared task, CCMatrix and CCAligned. We report the list of 12 language directions and the number of sentences (Millions) per each language pair.

References

Aharoni, R., Johnson, M., & Firat, O. (2019, June). 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)* (pp. 3874–3884). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N19-1388 doi: 10.18653/v1/N19-1388

Akhbardeh, F., Arkhangorodsky, A., Biesialska, M., Bojar, O., Chatterjee, R., Chaudhary, V., ... Zampieri, M. (2021, November). Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the sixth conference on machine translation* (pp. 1–88). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.wmt-1.1

Arivazhagan, N., Bapna, A., Firat, O., Aharoni, R., Johnson, M., & Macherey, W. (2019). *The missing ingredient in zero-shot neural machine translation.* arXiv. Retrieved from https://arxiv.org/abs/1903.07091 doi: 10.48550/ARXIV.1903.07091

Bapna, A., & Firat, O. (2019, November). Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 1538–1548). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1165 doi: 10.18653/v1/D19-1165

El-Kishky, A., Chaudhary, V., Guzman, F., & Koehn, P. (2019). *Ccaligned: A massive collection of cross-lingual web-document pairs.* arXiv. Retrieved from https://arxiv.org/abs/1911.06154 doi: 10.48550/ARXIV.1911.06154

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., ... Joulin, A. (2020). *Beyond english-centric multilingual machine translation.* arXiv. Retrieved from https://arxiv.org/abs/2010.11125 doi: 10.48550/ARXIV.2010.11125

Freitag, M., & Firat, O. (2020). *Complete multilingual neural machine translation.* arXiv. Retrieved from https://arxiv.org/abs/2010.10239 doi: 10.48550/ARXIV.2010.10239

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., ... Fan, A. (2022). The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, *10*, 522–538. Retrieved from https://aclanthology.org/2022.tacl-1.30 doi: 10.1162/tacl_a_00474

Ha, T.-L., Niehues, J., & Waibel, A. (2016, December 8-9). Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the 13th international conference on spoken language translation.* Seattle, Washington D.C: International Workshop on Spoken Language Translation. Retrieved from https://aclanthology.org/2016.iwslt-1.6

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., ... Zhou, M. (2018). *Achieving human parity on automatic chinese to english news translation.*

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., ... Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, *5*, 339–351. Retrieved from https://aclanthology.org/Q17-1024 doi: 10.1162/tacl_a_00065

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017, April). Bag of tricks for efficient text classification. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: Volume 2, short papers* (pp. 427–431).

Valencia, Spain: Association for Computational Linguistics. Retrieved from https://aclanthology.org/E17-2068

Kasai, J., Pappas, N., Peng, H., Cross, J., & Smith, N. A. (2020). Deep encoder, shallow decoder: Reevaluating non-autoregressive machine translation. *arXiv preprint arXiv:2006.10369*.

Kim, Y. J., Junczys-Dowmunt, M., Hassan, H., Aji, A. F., Heafield, K., Grundkiewicz, R., & Bogoychev, N. (2019). From research to production and back: Ludicrously fast neural machine translation. In *Proceedings of the 3rd workshop on neural generation and translation* (pp. 280–288).

Koehn, P. (2004, July). Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 388–395). Barcelona, Spain: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W04-3250

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., … Herbst, E. (2007, June). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions* (pp. 177–180). Prague, Czech Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P07-2045

Kudo, T., & Richardson, J. (2018, November). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 conference on empirical methods in natural language processing: System demonstrations* (pp. 66–71). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D18-2012 doi: 10.18653/v1/D18-2012

Lin, Z., Wu, L., Wang, M., & Li, L. (2021, August). Learning language specific sub-network for multilingual 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)* (pp. 293–305). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.acl-long.25 doi: 10.18653/v1/2021.acl-long.25

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., & Han, J. (2019). *On the variance of the adaptive learning rate and beyond.* arXiv. Retrieved from https://arxiv.org/abs/1908.03265 doi: 10.48550/ARXIV.1908.03265

Post, M. (2018, October). A call for clarity in reporting BLEU scores. In *Proceedings of the third conference on machine translation: Research papers* (pp. 186–191). Belgium, Brussels: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W18-6319

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020, November). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 2685–2702). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-main.213 doi: 10.18653/v1/2020.emnlp-main.213

Rios, A., Müller, M., & Sennrich, R. (2020, November). Subword segmentation and a single bridge language affect zero-shot neural machine translation. In *Proceedings of the fifth conference on machine translation* (pp. 528–537). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.wmt-1.64

Schwenk, H., Wenzek, G., Edunov, S., Grave, E., & Joulin, A. (2019). *Ccmatrix: Mining billions of high-quality parallel sentences on the web.* arXiv. Retrieved from https://

arxiv.org/abs/1911.04944 doi: 10.48550/ARXIV.1911.04944

Tiedemann, J. (2012, May). Parallel data, tools and interfaces in OPUS. In *Proceedings of the eighth international conference on language resources and evaluation (LREC'12)* (pp. 2214–2218). Istanbul, Turkey: European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf

Tran, C., Bhosale, S., Cross, J., Koehn, P., Edunov, S., & Fan, A. (2021). *Facebook ai wmt21 news translation task submission.* arXiv. Retrieved from https://arxiv.org/abs/2108.03265 doi: 10.48550/ARXIV.2108.03265

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). *Attention is all you need.*

Wang, X., Tsvetkov, Y., & Neubig, G. (2020, July). Balancing training for multilingual neural machine translation. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8526–8537). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.754 doi: 10.18653/v1/2020.acl-main.754

Wang, Y., Zhang, J., Zhai, F., Xu, J., & Zong, C. (2018, October-November). Three strategies to improve one-to-many multilingual translation. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2955–2960). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D18-1326 doi: 10.18653/v1/D18-1326

Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., … Liu, T. (2020). On layer normalization in the transformer architecture. In *International conference on machine learning* (pp. 10524–10533).

Yang, Y., Eriguchi, A., Muzio, A., Tadepalli, P., Lee, S., & Hassan, H. (2021). *Improving multilingual translation by representation and gradient regularization.* arXiv. Retrieved from https://arxiv.org/abs/2109.04778 doi: 10.48550/ARXIV.2109.04778

Zhang, B., Williams, P., Titov, I., & Sennrich, R. (2020). *Improving massively multilingual neural machine translation and zero-shot translation.* arXiv. Retrieved from https://arxiv.org/abs/2004.11867 doi: 10.48550/ARXIV.2004.11867