Zero-Shot Neural Machine Translation with Self-Learning Cycle

Surafel M. Lakew[†] **Matteo Negri** **Marco Turchi**
[†]University of Trento, Fondazione Bruno Kessler, Trento, Italy
{lakew,negri,turchi}@fbk.eu

Abstract

Neural Machine Translation (NMT) approaches employing monolingual data are showing steady improvements in resource-rich conditions. However, evaluations using real-world low-resource languages still result in unsatisfactory performance. This work proposes a novel *zero-shot* NMT modeling approach that learns without the now-standard assumption of a *pivot* language sharing parallel data with the zero-shot *source* and *target* languages. Our approach is based on three stages: *initialization* from any pre-trained NMT model observing at least the target language, *augmentation* of source sides leveraging target monolingual data, and *learning* to optimize the initial model to the zero-shot pair, where the latter two constitute a self-learning cycle. Empirical findings involving four diverse (in terms of a language family, script and relatedness) zero-shot pairs show the effectiveness of our approach with up to +5.93 BLEU improvement against a supervised bilingual baseline. Compared to unsupervised NMT, consistent improvements are observed even in a domain-mismatch setting, attesting to the usability of our method.

1 Introduction

Since the introduction of NMT (Sutskever et al., 2014; Bahdanau et al., 2014), model learning using unlabeled (*monolingual*) data is increasingly gaining ground. Undoubtedly, the main motivating factor to explore beyond *supervised* learning is the lack of enough (*parallel*) examples, a performance bottleneck regardless of the underlying architecture (Koehn and Knowles, 2017). A fairly successful approach using monolingual data is the *semi-supervised* learning with back-translation (Sennrich et al., 2015), particularly if the initial supervised model is good enough for augmenting quality pseudo-bitext (Poncelas et al., 2018; Ott et al., 2018; Caswell et al., 2019). Moreover, back-translation showed to be a core element of new monolingual based approaches. These include zero-shot NMT (Lakew et al., 2017; Gu et al., 2019; Currey and Heafield, 2019), which relies on a multilingual model (Johnson et al., 2017; Ha et al., 2016) (Fig. 1b) and unsupervised NMT, which initializes from pre-trained embeddings (Lample et al., 2018; Artetxe et al., 2018) or cross-lingual language model (Lample and Conneau, 2019) (Fig. 1d). At least two observations can be made on the approaches that leverage monolingual data: *i*) they require high-quality and comparable monolingual examples, and *ii*) they show poor performance on real-world zero-resource language pairs (ZRPs)[1] (Neubig and Hu, 2018; Guzmán et al., 2019).

To overcome these problems, in this work we propose a zero-shot modeling approach (Fig. 1c) to translate from an unseen *source* language U to a *target* language T that has only been

[†]Work conducted when the author was at FBK.
[1]ZRP: a language pair with only monolingual data available, alternatively called Zero-Shot Pair (ZSP).

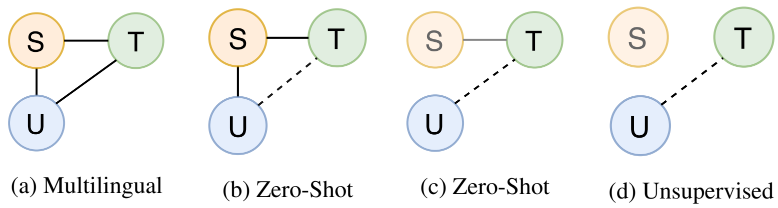

Figure 1: Proposed Zero-Shot (*c*) and existing NMT modeling approaches, using parallel (*solid line*) and monolingual (*broken line*) data.

observed by a model pre-trained on (S, T) parallel data (S being a different source language). In literature, zero-shot NMT has been investigated under the assumption that (U, P) and (P, T) parallel data involving a *pivot* language P are available for pre-training (Fig. 1b) (Johnson et al., 2017; Ha et al., 2016). However, most of the +$7,000$ currently spoken languages do not exhibit any parallel data with a common P language. *This calls for new techniques to achieve zero-shot NMT by reducing the requirements of the pivoting-based method.*

To this aim, our approach follows a self-learning cycle, by first translating in the *primal* zero-shot direction $U \rightarrow T$ with a model pre-trained on (*S, T*) parallel data that has never seen U during training. Then, the generated translations are used as a pseudo-bitext to learn a model for the *dual* $T \rightarrow U$ translation direction. This inference-learning cycle is iterated, alternating the dual and primal zero-shot directions until convergence of the $U \leftrightarrow T$ zero-shot model.

Through experiments on data covering eight language directions, we demonstrate the effectiveness of our approach in the ZRP scenario. In particular, we report significant improvements compared to both a *supervised* model trained in low-resource conditions (up to $+5.93$ BLEU) and an *unsupervised* one exploiting large multilingual corpora (up to $+5.23$ BLEU).
Our contributions can be summarized as follows:

- We propose a new variant of zero-shot NMT, which reduces the requirements of previous pivoting-based methods. Our approach enables incorporating an unseen zero-resource language U, with no need of pre-training on parallel data involving U.
- We empirically evaluate our approach on diverse language directions and in a real-world zero-resource scenario, a testing condition disregarded in previous literature.
- We provide a rigorous comparison against unsupervised neural machine translation, by testing our models in an in-domain, out-of-domain, and source to target domain mismatch scenarios.

2 Zero-Shot Translation

From a broad perspective, ZST research is moving in three directions, (*i*) improving translation quality by employing ZST specific objectives (Chen et al., 2017; Lu et al., 2018; Blackwood et al., 2018; Al-Shedivat and Parikh, 2019; Arivazhagan et al., 2019a; Pham et al., 2019; Ji et al., 2019; Siddhant et al., 2020), (*ii*) training favorable large scale multilingual models for the ZST languages with lexically and linguistically similar languages (Aharoni et al., 2019; Arivazhagan et al., 2019b), and (*iii*) incrementally learning better model for the ZST directions with self-learning objectives (Lakew et al., 2017; Gu et al., 2019; Zhang et al., 2020). The common way of employing *self-supervised learning* in ZST modeling is iterative back-translation that generates the source from the monolingual target to construct a new parallel sentence pair. In terms of performance, while (*i*) and (*ii*) fall behind, (*iii*) either approaches or even outperforms the two-step *pivot translation* approach ($S \rightarrow P \rightarrow T$).

Despite these progresses, current approaches make identical assumptions, namely: *i) the* reliance on a multilingual model, and *ii)* observing both the U and T zero-shot languages paired with the P language(s). However, in a real-world setting these conditions are rarely satisfied, hindering the application of zero-shot NMT to the vast majority of ZRP. Moreover, conditioning ZST on P language(s) creates a performance ceiling that depends on the amount, domain, and quality of parallel data available for the $S - P$ and $P - T$ pairs.

To our knowledge, a zero-shot NMT modeling between an *unseen*-U and T zero-shot pair, and without the *P* language(s) criterion has not yet been explored, motivating this work.

2.1 Zero-Shot Translation with Self-Learning

We propose a new zero-shot NMT (zNMT) approach that expands the current definition of zero-shot to the extreme scenario, where we avoid the established assumption of observing the zero-shot (U, T) languages paired with the pivot (P) language(s). Instead, we consider only the availability of monolingual data for (U, T) and a pre-trained NMT observing only the T language $-$ a scenario applicable to most zero-resource languages.

To this end, with the goal of learning a zero-shot model covering the primal ($U \rightarrow T$) and the dual ($T \rightarrow U$) directions, our zNMT approach consists of three stages: model initialization, incremental data augmentation, and model learning. The latter two steps can be iterated over time creating a *self-learning cycle* between the primal and dual zero-shot directions.

2.2 Model Initialization

Different from conventional ZST approaches, zNMT can be either initialized from a bilingual or a multilingual pre-trained system (Algorithm 1, *line* 2). The only assumption we consider is the availability of the zero-shot T side at pre-training time. Hence, our initialization introduces relaxation to model pre-training, a direct consequence of removing the P language premise. Considering that U has never been observed, we analyze different pre-training strategies to build a robust zero-shot system.

Algorithm 1: Proposed zNMT

1 Input;
$U_m, T_m \leftarrow$ monolingual data of the ZSP;
TM $\leftarrow$ pre-trained translation model;
$R \leftarrow$ maximum self-learning cyle;
2 zNMT^0, $\leftarrow$ TM;
3 $D_{Pr} = \emptyset$; $D_{Du} = \emptyset$;
4 $r \leftarrow 1$;
5 **for** r **to** R **do**
6 $\quad T^* \leftarrow Primal\_Infer$ (zNMT^0, U_m)
7 $\quad D_{Du} \leftarrow (T^*, U_m)$
8 $\quad$ zNMT$^1 \leftarrow$ Train (zNMT^0, $D_{Du} \cup D_{Pr}$)
9 $\quad D_{Pr} = \emptyset$
10 $\quad U^* \leftarrow Dual\_Infer$ (zNMT^1, T_m)
11 $\quad D_{Pr} \leftarrow (U^*, T_m)$
12 $\quad$ zNMT$^2 \leftarrow$ Train (zNMT^1, $D_{Pr} \cup D_{Du}$)
13 $\quad$ zNMT$^0 \leftarrow$ zNMT^2
14 $\quad D_{Du} = \emptyset$
end
15 **return** zNMT^0

2.3 Self-Learning Cycle

In our scenario, we have only access to monolingual data for both the U and T languages, and the pre-trained model did not leverage $U - T$ parallel data during training. In this setting, after the initialization, the very first task is a zero-shot inference for $U \rightarrow T$ (*line* 6), to which we refer as the *primal* ZST direction. The goal of this step is to acquire pseudo-bitexts to enhance the translation capability of the NMT system for the $T \rightarrow U$ direction.

To this aim, when the primal inference is concluded, a learning step is performed by reverting the generated pseudo-bitext data ($T \rightarrow U$). zNMT optimizes the same objective function (Eq. 1) as in the pre-training or Eq. 2 if zero-shot training is performed together with other supervised languages pairs (*co-learning*) as in (Johnson et al., 2017). The resulting model is then used to perform the $T \rightarrow U$ inference (*line* 10), to which we refer as the *dual* direction. The data generated by the dual process is then paired with the data produced by the primal one and used in a new leaning step. Assuming that at each inference step our algorithm generates better quality bi-texts, we replace the dual or primal data (D_{Du} and D_{Pr}) produced at the previous round with the ones generated in the current round. For instance, during the training at line 8, we use the D_{Du} generated at line 7, while we keep the D_{Pr} generated at line 11 of the previous round.

This sequential approach alternates the primal and dual inferences (*lines* 6, 10) with a learning phase in between (*lines* 8, 12). The goal of this procedure is two-fold, first to implement a self-learning strategy and then to acquire more and better pseudo-bitext pairs.

Unlike previous work on improving zero-shot translation (Lakew et al., 2017; Gu et al., 2019), we focus on learning only a model for the ZSP languages. However, for a better analysis and fair comparison with multilingual supervised approaches, we further show ZNMT performance by co-learning with language pairs with parallel data (*such as incorporating the $S-T$ pair while learning the zero-shot $U-T$*).

An important aspect for ZNMT is how close is U to S in terms of vocabulary and sentence structure. Our intuition is that the closer the two languages, the higher is the performance achieved by the ZNMT. This will be explored in §4.

3 Experimental Settings

To build an experimental ground that defines zero-shot translation without the pivot language, we selected a real-world low-resource languages benchmark. In other words, we considered the data to incorporate multiple and diverse languages, including parallel data for building strong baselines and monolingual data to evaluate ZNMT. Moreover, our choice is motivated by the findings of Neubig and Hu (2018) and Guzmán et al. (2019), showing that monolingual-based approaches under-perform when assessed with real-world zero-shot pairs (ZSPs).

3.1 Languages and Dataset

Due to their low-resource nature, we use Ted talks data (Cettolo et al., 2012; Qi et al., 2018) for Azerbaijani (Az), Belarussian (Be), Galician (Gl), and Slovak (Sk) paired with English (En). The four pairs come with train, dev, and test sets, with a max of $61k$ and as few as $4.5k$ examples, creating an ideal scenario of low-resource pair (LRP). The parallel data of the LRP is used to build baseline models in isolation and in a multilingual settings. The same dataset has been also used in recent works in an extremely low-resource scenario (Neubig and Hu, 2018; Xia et al., 2019; Lakew et al., 2019).

For the approaches utilizing monolingual data, we take the non–En side for each of the four LRP languages as in-domain (IND) monolingual examples. For the En monolingual data, segments are collected from the target side of the respective $S-T(En)$ pairs. However, to avoid the presence of comparable sentences in the U and T sides of the ZSP, we discard segments of monolingual En if the $T(En)$ side of the $U-T$ and $S-T$ are overlapping.

For out-of-domain (OOD) monolingual data we extract segments from Wikipedia dumps, similar to Xia et al. (2019).[2] The collected data are de-duplicated and overlapping segments with the IND monolingual are removed. To create a practical real-world scenario that represents most of the ZRP languages, we take only the top 2×10^6 segments, aligning with the maximum number of samples that are available for the non-En languages in this benchmark. Statistics about the data are shown in Table 1.

[2]Wikipedia: https://dumps.wikimedia.org/

		Domain	Az-En	Be-En	Gl-En	Sk-En
Sample Size	Parallel (Train/Dev/Test)	IND	5.9k	4.5K	10.0k	61.5k
		IND	671	248	682	2271
		IND	903	664	1007	2445
	Monolingual	IND	5.9k, 174k	4.5k, 201k	10K, 174K	61.5k, 58.6
		OOD	1.85M, 2M	1.67M, 2M	1.9M, 2M	1.8M, 2M
U Language Property	Family/Script		Turkic/Latin	Slavic/Cyrillic	Romance/Latin	Slavic/Latin

Table 1: Languages and data statistics for parallel in-domain (IND) LRP, monolingual IND and out-domain (OOD).

3.2 Models

To test the ZNMT strategy and compare it against other approaches, we train the following models:

- **NMT**: trained with *supervised* objective using parallel IND data of each LRP.
- **sNMT** (semi-supervised NMT): trained with *semi-supervised* objective (Sennrich et al., 2015) with back-translation leveraging OOD monolingual data.
- **MNMT** (multilingual NMT): trained with *supervised* objective aggregating 116 directions IND parallel data (Johnson et al., 2017).[3]
- **uNMT** (unsupervised NMT): trained with *unsupervised* objectives (Lample and Conneau, 2019) leveraging IND and OOD data.
- **zNMT** (zeros-shot NMT): trained with proposed *zero-shot* modeling leveraging IND and OOD monolingual data.

3.3 Pre-Training Objectives

UNMT leverages a cross-lingual masked language model pre-training (MLM). We train the MLM following the (Lample and Conneau, 2019) settings, using both OOD and IND monolingual data of each ZSP language. Although ZNMT can be initialized from any pre-trained NMT model as long as the T language of the ZSP is observed (see §2.1), we devise three types of pre-training strategies for a rigorous evaluation and based on data availability:

- BITM – four bilingual translation models trained with $S \leftrightarrow En$ parallel data.[4]
- MUTM100 – a multilingual NMT model with 100 translation directions from the TED talks data, excluding the four ZSP and the pairs used for BITM.
- MUTM108 – a similar multilingual model with MUTM100, however, including additional 8 directions used for the BITM models.

The idea behind the MUTM100 and MUTM108 strategies is to check to what extent the presence of close languages to the $unseen\text{-}U$ in the pre-trained model can support the ZNMT approach. Note that, unlike in the MLM, all the pre-training for ZNMT utilized only in-domain[5]

[3] List of languages can be found in the Appendix.

[4] S is a related language to the $unseen\text{-}U$. The $S/unseen\text{-}U$ combinations are: Az/Turkish(Tr), Be/Russian(Ru), Gl/Portuguese(Pt), and Sk/Czech(Cs).

[5] Utilizing OOD monolingual data for NMT pre-training could be an advantageous and interesting direction to investigate, however, for this work we constrain to utilizing only IND data.

Id	Model	Pre-Train	Scen.	Az-En	En-Az	Be-En	En-Be	Gl-En	En-Gl	Sk-En	En-Sk
1	Supervised (NMT)	-	IND	3.60	2.07	5.20	3.40	19.53	15.52	27.24	20.91
2	Semi-Supervised (SNMT)	-	IOD	3.74	1.92	5.74	4.03	22.08	17.27	**27.85**	**21.24**
3	Unsupervised (UNMT)	MLM	IND	1.97	1.56	4.61	1.47	13.93	5.89	15.70	11.91
4			OOD	3.26	2.55	5.69	3.73	16.71	14.90	10.62	7.62
5			I-OD	0.88	1.18	0.82	0.90	5.06	2.78	6.39	7.28
6			IOD	3.97	2.57	5.57	3.78	20.23	17.07	13.77	11.43
7	Zero-Shot (ZNMT)	BITM	IND	8.86	4.87	4.42	3.45	23.57	18.17	17.89	14.08
8			OOD	6.76	4.45	5.75	5.16	17.28	16.97	9.13	6.74
9			I-OD	2.63	3.96	1.20	2.23	14.96	16.23	9.10	11.35
10			IOD	**11.38**	**6.28**	**7.36**	**6.35**	**25.46**	**21.09**	19.43	14.70

Table 2: Results from low-resource supervised and semi-supervised, and our monolingual based ZNMT in comparison with UNMT (Lample and Conneau, 2019) across the four training scenarios.

3.4 Training Scenarios

We define four model training criteria based on a real-world scenario for a ZSP, that is the availability and characteristics (*such as domain and size*) of monolingual data.

- **IND**: in-domain data is used both on the U and T zero-shot sides.
- **OOD**: out-of-domain data are used both in the U and T sides of the ZSP.
- **I-OD**: a scenario where we create a domain mismatch between the U and T side of the ZSP, by replacing the T IND with OOD data.
- **IOD**: a the mix of IND and relatively large OOD data is used on both U and T sides.

3.5 Training Pipeline

Data Preparation: we collect the IND Ted talks data provided by Qi et al. (2018) and OOD Wikipedia[6] data, and then segment them into sub-word units. We use SentencePiece (Kudo and Richardson, 2018)[7] to learn BPE with 32k merge operations using the IND training data, whereas for UNMT we also use OOD monolingual data.

Model Settings: all experiments use Transformer (Vaswani et al., 2017). UNMT is trained using the XLM tool (Lample and Conneau, 2019)[8], while for the rest we utilize OpenNMT (Klein et al., 2017).[9] Models are configured with 512 dimension, 8 headed 6 self-attention layers, and 2048 feed-forward dimension. Additional configuration details are provided in the Appendix.

Evaluation: we use the BLEU score (Papineni et al., 2002)[10] for assessing models' performance. Scores are computed on detokenized (*hypothesis*, *reference*) pairs. The checkpoints with best BLEU on the dev set are used for the final evaluations.

4 Results and Analysis

In Table 2, we asses the quality of various NMT systems featuring different model types (§3.2), training scenarios (§3.4) using bilingual BITM for ZNMT and MLM for UNMT pre-training.

[6]WikiExtractor: https://github.com/attardi/wikiextractor
[7]SentencePiece: https://github.com/google/sentencepiece
[8]XLM: https://github.com/facebookresearch/XLM
[9]OpenNMT: https://github.com/OpenNMT
[10]Moses Toolkit: http://www.statmt.org/moses

Id	Model	Pre-Train	Scen.	Az-En	En-Az	Be-En	En-Be	Gl-En	En-Gl	Sk-En	En-Sk
1	Supervised (MNMT)	-	IND	**11.37**	**4.98**	**18.36**	10.06	29.77	25.44	27.49	**22.72**
2	Zero-Shot (ZNMT)	MUTM100	I-OD	1.04	2.55	7.31	7.14	22.91	22.93	12.33	11.81
3			IOD	2.51	1.55	16.20	10.30	32.14	26.68	23.52	16.60
4		MUTM108	I-OD	4.14	2.38	10.18	9.00	26.45	25.34	20.26	19.69
5			IOD	9.19	2.75	17.26	**10.95**	**32.83**	**27.49**	**28.94**	21.53

Table 3: ZNMT when initialized from multilingual pre-trained models, in comparison with supervised MNMT.

We then show the effect of leveraging massive multilingual pre-training on the ZNMT performance (Table 3). Finally, we expand our analysis to co-learning ZNMT with supervised NMT (Table 4). A preliminary assessment of the experimental choices adopted for ZNMT can be found in the Appendix.

4.1 Bilingual Pre-Training

The first two rows of Table 2 confirm the results of (Sennrich et al., 2015) showing that semi-supervised approaches, which leverage back-translation, outperform supervised NMT systems. Moreover, the performance of both approaches strongly relates to the quantity of the available training data ($Gl - En \gg Be - En$).

In the **in-domain** training scenario (IND), our ZNMT approach outperforms the supervised low-resource NMT, except for $Sk \leftrightarrow En$ and $Be \rightarrow En$ ($rows$ 1, 7), demonstrating the effectiveness of our proposal in leveraging monolingual data. The advantage of ZNMT is further confirmed when comparing it with UNMT. In this case, ZNMT outperforms UNMT in 7 out of 8 language directions and it is on par on the $Be \rightarrow En$ language pair.

In the **out-of-domain** training scenario (OOD), despite the fact that UNMT utilizes $\times 10$ more OOD segments than ZNMT, our approach surprisingly achieves better performance than UNMT, except for $Sk - En$ ($rows$ 4, 8). Fig. 2 shows the effect of varying the amount of monolingual data during pre-training (BITM, MLM). We observe that UNMT is significantly affected by decreasing the size of the monolingual data and, when using the same quantity applied in ZNMT (200k), it achieves much worse performance (-10 BLEU points). Our findings clearly show the effectiveness of ZNMT in learning better with small monolingual data, a case applicable for most LRP and ZSP.

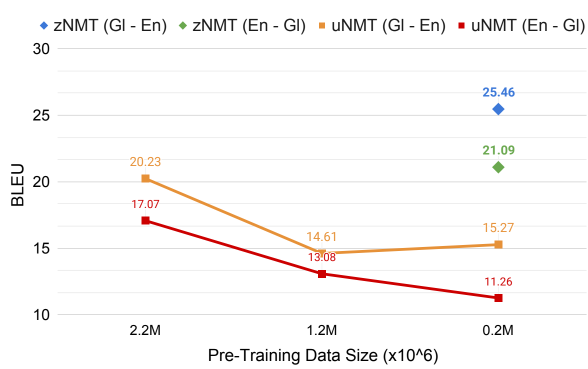

Figure 2: Effect of pre-training data size.

The **domain mismatch** scenario (I-OD) is the most realistic representation for ZSP and LRP settings, as it does not count on access to comparable monolingual data. Both ZNMT and UNMT show drastic performance drops in all directions (*rows* 5, 9), confirming the findings of Kim et al. (2020). Besides the performance drop, ZNMT shows higher robustness to domain shifts, resulting in higher scores. UNMT, in contrast, is susceptible to the domain divergence and requires comparable monolingual data that is hard to acquire for ZRP.

In the **mixed domain** scenario (IOD), ZNMT prevails over UNMT by a larger margin

Id	Model	Pre-Train	Scen.	Az-En	En-Az	Be-En	En-Be	Gl-En	En-Gl	Sk-En	En-Sk
1	MASSIVE	-	IND	12.78	5.06	21.73	**10.72**	30.65	26.59	29.54	24.52
2	DYNADAPT	MUTM108	IND	15.33	-	23.80	-	**34.18**	-	**32.48**	-
3	AUGADAPT	MUTM108	IOD	**15.74**	-	**24.51**	-	33.16	-	32.17	-
4	ZNMT + CO-LEARNING	MUTM108	IND	8.56	2.25	16.34	9.16	32.75	26.75	28.84	22.34
5			IOD	11.01	2.28	18.05	10.38	33.26	**27.64**	29.94	22.26

Table 4: Performance of ZNMT with co-Learning in comparison with the supervised MASSIVE (Aharoni et al., 2019) and DYNADAPT (Lakew et al., 2019), and semi-supervised AUGADAPT (Xia et al., 2019).

(*rows* 6, 10) ranging from $+1.79$ ($Az - En$) to +7.41 ($Be - En$) BLEU. This is a similar trend to the IND, OOD, and I-OD scenarios, validating the superiority of the proposed ZNMT learning approach. A more interesting aspect is that, except for $Sk - En$ (*rows* 2, 10), ZNMT also outperforms semi-supervised NMT. This shows that a stronger model can be learned exploiting as few as $200k$ monolingual data with our ZNMT learning principles, in comparison with a LRP performance (such as $Az - En$ with $5.9k$, and $Gl - En$ with 10k parallel data).

In sum, the results in Table 2 show that, for low-resource language pairs, ZNMT leveraging BITM can in most of the cases outperform supervised NMT trained on language-specific parallel data. Moreover, ZNMT is robust towards domain shifts from the pre-training and across $U - T$ ZSP, outperforming unsupervised NMT in all training scenarios.

4.2 Multilingual Pre-Training

To test the capability of ZNMT to leverage universal representation from the pre-trained model, we built two massive multilingual systems: MUTM100 that excludes the pairs used for BITM and MUTM108 that assumes a favorable condition by also including the BITM pairs.

Besides the initialization from the universal models, we train the best ZNMT scenario (IOD) and the most challenging one (I-OD) from Table 2 by only using the $U - T$ monolingual data. Table 3 (*row* 1) shows that the supervised MNMT benefits from the multilingual corpus (i.e., trained with 116 directions data including the zero-shot pairs), and, as expected, obtains improvements over the bilingual supervised models in Table 2.

In the **domain mismatch** scenario (I-OD), the use of BITM leads to large drops in performance compared to the IND or IOD scenario (Table 2, *row* 9). This is also confirmed when leveraging the MUTM* pre-training (*rows* 2, 4). However, the robust multilingual pre-training shows improvements compared with the initialization from BITM. For instance, the $Gl \rightarrow En$ with BITM drops -10.5 (from 25.46 with IOD to 14.96 with I-OD), while MUTM108 degrades only by 6.38 BLEU points.

Our approach leveraging the **mixed domain** (IOD) monolingual data with MUTM108 achieves the best performance in most of the language directions and is on par or even better with the supervised multilingual (*rows* 1, 3, and 5). This is a remarkable result because the ZNMT systems do not leverage any language-specific parallel data.

The advantage of using a **robust pre-training** can be ascribed to the availability of multiple languages that maximizes the lexical and linguistic similarity with the ZSP. Looking at the IOD scenario MUTM* in Table 3 (*rows* 3, 5), we notice an overall improvement over BITM pre-training (Table 2, *row* 10). A comparison against the best supervised SNMT model (Table 2) using low-resource parallel data shows better performance of ZNMT with MUTM108 up to (+10.75 $\leftrightarrow$ +10.22) for $Gl - En$. However, as for the BITM, it is not always the case to find closely related $S - T$ pair(s) to the $U - T$ ZSP for pre-training. Hence, it is rather more interesting to observe that ZNMT can learn even better with MUTM100 without observing the most related languages as in BITM.

With respect to the BITM, $Az-En$ is the only ZSP that do not benefit from the multilingual pre-training. One possible reason is the absence of related language pairs, which makes the pre-training representation dominated by other pairs. This becomes more evident (lower BLEU) when using MUTM100, a pre-training that excludes the closest $S-T$ pair to $Az-En$.

In sum, our approach shows significant improvements when leveraging universal pre-trained models. This is demonstrated by the large gains in performance in all the scenarios over the BITM pre-training. The fact that our method is able to approach the performance of the multilingual supervised settings, and in some cases to overcome them, makes it a valuable solution for ZSP languages.

4.3 Co-Learning with Supervised Directions

To test the complementary of ZNMT and supervised NMT, we add the parallel data of the latter only at the learning stage. Although it is possible to leverage multilingual parallel data, in this experiments we only utilize a single $S-T$ parallel pair from the BITM for the zero-shot co-learning stage of $U-T$.

We compare our co-learning system with three state-of-the-art approaches: MASSIVE trains a many-to-many system on all ($116 \leftrightarrow 116$) available pairs (Aharoni et al., 2019), DYNADAPT (Lakew et al., 2019) uses an IND criterion to adapt MUTM108 pre-trained model by first tailoring the vocabulary and embeddings to the LRP and AUGADAPT (Xia et al., 2019) generates pseudo-bitext from OOD monolingual and adapts MUTM108 together with the IOD data. The latter two utilize a similar co-learning strategy during the adaptation of the universal model with the parallel data, and reported results only when the target is En. Similar to MASSIVE, DYNADAPT, and AUGADAPT, we focused on IND and IOD training scenarios.

Table 4 reports the performance of these approaches and of ZNMT with co-learning using in the IND and IOD scenarios. Comparing ZNMT + CO-LEARNING (*rows* 5) with ZNMT in Table 3 (*row* 5), the results show that *co-learning* generally leads to better performance. However, when the target language is non-En, the differences are marginal and the two approaches can be considered comparable. This is directly associated with the fact that we have more En segments, from the aggregation of the $S-T(En)$ and $U-T(En)$ pairs. DYNADAPT and AUGADAPT are the two best performing supervised techniques on the this benchmark, but ZNMT with co-learning achieves competitive performance approaching them both in the IND and IOD scenarios.

Overall, these findings show that our approach makes it possible to extend zero-shot NMT to an unseen language U. In particular, leveraging a universal pre-training model and co-learning with supervised task allows our approach to learn a better NMT model from monolingual data.

5 Conclusion

We presented a new zero-shot NMT modeling variant, specifically targeting languages that have never been observed in a pre-trained NMT. We showed limitations of current approaches with the pivot language premise and zero-shot translation only between observed languages, and proposed a relaxation to zero-shot NMT to incorporate unseen languages. Our approach includes initialization, augmentation, and training stages to construct a self-learning cycle to incrementally correct the primal and dual zero-shot translation quality. We empirically demonstrated the effectiveness of the proposed approach using diverse real-world zero-resource languages in in-domain, out-of-domain, domain-mismatch, and mixed domain scenarios. Results both from bilingual and multilingual initialization not only revealed the possibility of extending zero-shot NMT for unseen languages but also improved performance over unsupervised, low-resource supervised and semi-supervised NMT.

References

Aharoni, R., Johnson, M., and Firat, O. (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.

Al-Shedivat, M. and Parikh, A. P. (2019). Consistency by agreement in zero-shot neural machine translation. *arXiv preprint arXiv:1904.02338*.

Arivazhagan, N., Bapna, A., Firat, O., Aharoni, R., Johnson, M., and Macherey, W. (2019a). The missing ingredient in zero-shot neural machine translation. *arXiv preprint arXiv:1903.07091*.

Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M. X., Cao, Y., Foster, G., Cherry, C., et al. (2019b). Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2018). Unsupervised neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations*.

Artetxe, M., Ruder, S., Yogatama, D., Labaka, G., and Agirre, E. (2020). A call for more rigor in unsupervised cross-lingual learning. *arXiv preprint arXiv:2004.14958*.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bertoldi, N. and Federico, M. (2009). Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the fourth workshop on statistical machine translation*, pages 182–189. Association for Computational Linguistics.

Blackwood, G., Ballesteros, M., and Ward, T. (2018). Multilingual neural machine translation with task-specific attention. *arXiv preprint arXiv:1806.03280*.

Bojar, O. and Tamchyna, A. (2011). Improving translation model by monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336. Association for Computational Linguistics.

Caswell, I., Chelba, C., and Grangier, D. (2019). Tagged back-translation. *arXiv preprint arXiv:1906.06442*.

Cettolo, M., Girardi, C., and Federico, M. (2012). Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, volume 261, page 268.

Chen, Y., Liu, Y., Cheng, Y., and Li, V. O. (2017). A teacher-student framework for zero-resource neural machine translation. *arXiv preprint arXiv:1705.00753*.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Currey, A. and Heafield, K. (2019). Zero-resource neural machine translation with monolingual pivot data. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 99–107.

Dong, D., Wu, H., He, W., Yu, D., and Wang, H. (2015). Multi-task learning for multiple language translation. In *ACL (1)*, pages 1723–1732.

Firat, O., Cho, K., and Bengio, Y. (2016a). Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073*.

Firat, O., Sankaran, B., Al-Onaizan, Y., Vural, F. T. Y., and Cho, K. (2016b). Zero-resource translation with multi-lingual neural machine translation. *arXiv preprint arXiv:1606.04164*.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252. JMLR. org.

Gu, J., Wang, Y., Cho, K., and Li, V. O. (2019). Improved zero-shot neural machine translation via ignoring spurious correlations. *arXiv preprint arXiv:1906.01181*.

Guzmán, F., Chen, P.-J., Ott, M., Pino, J., Lample, G., Koehn, P., Chaudhary, V., and Ranzato, M. (2019). Two new evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. *arXiv preprint arXiv:1902.01382*.

Ha, T.-L., Niehues, J., and Waibel, A. (2016). Toward multilingual neural machine translation with universal encoder and decoder. *arXiv preprint arXiv:1611.04798*.

Ji, B., Zhang, Z., Duan, X., Zhang, M., Chen, B., and Luo, W. (2019). Cross-lingual pre-training based transfer for zero-shot neural machine translation. *arXiv preprint arXiv:1912.01214*.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistic*, 5:339–351.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709.

Kim, Y., Graça, M., and Ney, H. (2020). When and why is unsupervised neural machine translation useless? *arXiv preprint arXiv:2004.10581*.

Kim, Y., Petrov, P., Petrushkov, P., Khadivi, S., and Ney, H. (2019). Pivot-based transfer learning for neural machine translation between non-english languages. *arXiv preprint arXiv:1909.09524*.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Kocmi, T. and Bojar, O. (2018). Trivial transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1809.00357*.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Kudo, T. and Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Lakew, S. M., Erofeeva, A., Negri, M., Federico, M., and Turchi, M. (2018). "Transfer Learning in Multilingual Neural Machine Translation with Dynamic Vocabulary". In *15th International Workshop on Spoken Language Translation (IWSLT)*, Bruges, Belgium.

Lakew, S. M., Karakanta, A., Federico, M., Negri, M., and Turchi, M. (2019). "Adapting Multilingual Neural Machine Translation to Unseen Languages". In *16th International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong.

Lakew, S. M., Lotito, Q. F., Matteo, N., Marco, T., and Marcello, F. (2017). Improving zero-shot translation of low-resource languages. In *14th International Workshop on Spoken Language Translation*.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Lample, G., Denoyer, L., and Ranzato, M. (2018). Unsupervised machine translation using monolingual corpora only. In *Proceedings of the 6th International Conference on Learning Representations*.

Lu, Y., Keung, P., Ladhak, F., Bhardwaj, V., Zhang, S., and Sun, J. (2018). A neural interlingua for multilingual machine translation. *arXiv preprint arXiv:1804.08198*.

Luong, M.-T., Le, Q. V., Sutskever, I., Vinyals, O., and Kaiser, L. (2015). Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

Neubig, G. and Hu, J. (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. Association for Computational Linguistics.

Nguyen, T. Q. and Chiang, D. (2017). Transfer learning across low-resource, related languages for neural machine translation. *arXiv preprint arXiv:1708.09803*.

Ott, M., Edunov, S., Grangier, D., and Auli, M. (2018). Scaling neural machine translation. *arXiv preprint arXiv:1806.00187*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Pham, N.-Q., Niehues, J., Ha, T.-L., and Waibel, A. (2019). Improving zero-shot translation with language-independent constraints. *arXiv preprint arXiv:1906.08584*.

Poncelas, A., Shterionov, D., Way, A., Wenniger, G. M. d. B., and Passban, P. (2018). Investigating backtranslation in neural machine translation. *arXiv preprint arXiv:1804.06189*.

Press, O. and Wolf, L. (2016). Using the output embedding to improve language models. *arXiv preprint arXiv:1608.05859*.

Qi, Y., Sachan, D. S., Felix, M., Padmanabhan, S. J., and Neubig, G. (2018). When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of NAACL-HLT 2018*, pages 529–535. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2015). Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Sestorain, L., Ciaramita, M., Buck, C., and Hofmann, T. (2018). Zero-shot dual machine translation. *arXiv preprint arXiv:1805.10338*.

Siddhant, A., Bapna, A., Cao, Y., Firat, O., Chen, M., Kudugunta, S., Arivazhagan, N., and Wu, Y. (2020). Leveraging monolingual data with self-supervision for multilingual neural machine translation. *arXiv preprint arXiv:2005.04816*.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Xia, M., Kong, X., Anastasopoulos, A., and Neubig, G. (2019). Generalized data augmentation for low-resource translation. *arXiv preprint arXiv:1906.03785*.

Xia, Y., He, D., Qin, T., Wang, L., Yu, N., Liu, T., and Ma, W. (2016). Dual learning for machine translation. *CoRR*, abs/1611.00179.

Zhang, B., Williams, P., Titov, I., and Sennrich, R. (2020). Improving massively multilingual neural machine translation and zero-shot translation. *arXiv preprint arXiv:2004.11867*.

Zoph, B., Yuret, D., May, J., and Knight, K. (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. Association for Computational Linguistics.

Appendices

A Background and Motivation

For an S and T language pair, a standard NMT model is learned by mapping (s, t) example pairs with an encoder-decoder network (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014) such as Recurrent (Bahdanau et al., 2014; Cho et al., 2014), Convolutional (Gehring et al., 2017), and Transformer (Vaswani et al., 2017). Despite the varied architectural choices, the objective of NMT is to minimize the loss,

$$L_{\hat{\theta}}(s,t) = \sum_{i=1}^{|t|+1} logp(t_i|s,\hat{\theta}) \tag{1}$$

$\hat{\theta}$ is the parameterization of the network, s is the source sentence, and t is the predicted sentence. Reserved tokens $\langle bos \rangle$ at $i = 0$ and $\langle eos \rangle$ at $i = |t| + 1$ defines the beginning and end of t.

Training Paradigms

Semi-Supervised learning leverages monolingual data and has been used to improve supervised phrase-based models (Bertoldi and Federico, 2009; Bojar and Tamchyna, 2011). In NMT the procedure is commonly called – *back-translation* (Sennrich et al., 2015); to enhance $S \rightarrow T$ direction additional pseudo-bitext is utilized by augmenting the S side from T monolingual segments with a reverse $T \rightarrow S$ model. Back-translation became a core module in approaches that leverage monolingual data, such as dual-learning (Xia et al., 2016; Sestorain et al., 2018), zero-shot (Firat et al., 2016b; Lakew et al., 2017; Gu et al., 2019; Currey and Heafield, 2019; Zhang et al., 2020), and unsupervised (Lample et al., 2018; Artetxe et al., 2018) translation.

Unsupervised learning considers only monolingual data of S and T languages. Initialization from pre-trained embeddings (Artetxe et al., 2018; Lample et al., 2018) or cross-lingual language model (Lample and Conneau, 2019), denoising auto-encoder and iterative back-translation are commonly employed learning objectives. Despite being a rapidly growing research area, findings show failures in an unsupervised NMT when using real-world ZRP (Neubig and Hu, 2018), distant languages (Guzmán et al., 2019) in a domain-mismatched scenario (Kim et al., 2020; Artetxe et al., 2020). Given the similarity in leveraging monolingual data, we directly compare our zero-shot NMT with unsupervised NMT.

Multilingual modeling extends Eq. 1 objective to multiple language pairs. Although early work dedicates network components per language (Dong et al., 2015; Luong et al., 2015; Firat et al., 2016a), the most effective way utilizes "*language-id*" to share a single encoder-decoder model across multiple language pairs (Johnson et al., 2017; Ha et al., 2016). For L languages, a model learns to maximize the likelihood over all the available language pairs (max of $N = L(L - 1)$) parallel data. For each language pair $S, T \in N$ and $S \neq T$, Eq. 1 can be written as,

$$L_{\hat{\theta}}^{T}(s^S,t^T) = \sum_{i=1}^{|t|+1} logp(t_i^T|s^S,\hat{\theta}) \tag{2}$$

Where the *language-id* (i.e. $\langle 2T \rangle$) is explicitly inserted at i=1 of the source (s^T). Most importantly, multilingual modeling enables translation between language pairs without an actual training data ($S, T \notin N$), exploiting an implicit transfer-learning from pairs with training data – also known as *Zero-Shot Translation* (ZST) (Johnson et al., 2017).

Transfer Learning across NMT models (i.e, *parent-to-child*) (Zoph et al., 2016), have been shown to work effectively with a shared vocabulary across related (Nguyen and Chiang, 2017) and even distant (Kocmi and Bojar, 2018) languages, by pre-training multilingual models (Neubig and Hu, 2018), by updating parent vocabularies with child (Lakew et al., 2018), and for a ZST with a pivot language (Kim et al., 2019). In this work, we leverage for the first time pre-trained models for zero-shot translation without the pivot language assumption.

B Preliminary Assessment

We summarize the motivation for certain experimental design choices in our zero-shot NMT (ZNMT) modeling, analyzing model pre-training type (such as bilingual (BITM) and multilingual (MUTM*)) and effective utilization of the in-domain (IND), out-of-domain (OOD), mixed domain (IOD) monolingual data. The $Gl - En$ zero-shot pair (ZSP) is used for our assessment.

Pre-Trained NMT Variant

Unlike previous work in zero-shot NMT, our ZNMT aims to leverage both bilingual and multilingual pre-trainings. Fig. 3 shows ZNMT improves better if initialized from multilingual pre-training (MUTM100). This is despite MUTM100 not observing the closest language pair ($Pt - En$) to the ZSP ($Gl - En$), while BITM is trained using only $Pt - En$. Hence, the gain by initializing from MUTM100 shows the robustness of pre-training with multiple languages and its positive effect on ZNMT. However, these results signal MUTM* importance for ZNMT modeling, for further verification and better comparison with the bilingual supervised and unsupervised approaches our main experimental setup first focuses on utilizing BITM.

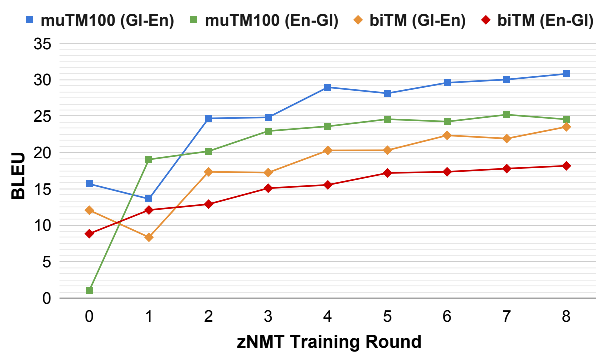

Figure 3: Performance of ZNMT using bilingual (BITM) and multilingual (MUTM100) pre-trainings.

Data Size and Domain

For the training scenarios involving IND and OOD data, Fig. 4 shows if available using all IND segments (All:All) is better than taking equal proportion (1:1) of the U and T sides of the ZSP. In a parallel experiment for IOD scenario, however, we observed that balancing the OOD segments with IND lead to a comparable or better performance. In other words, we select OOD proportionally ($\approx 200k$) to the largest IND side of the ZSP. We noted a similar trend for semi-supervised (SNMT) low-resource model, that shows better performance when using $\approx 200k$ OOD leading to

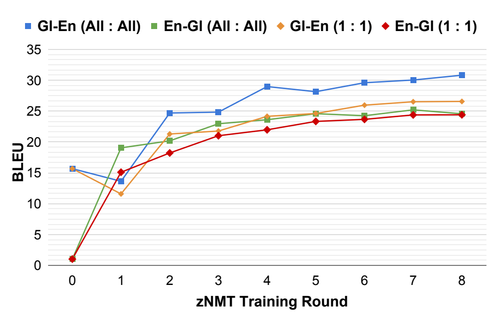

Figure 4: Performance of ZNMT by varying monolingual data size ratio between U and T.

22.08 $\leftrightarrow$ 17.27 ($Gl \leftrightarrow En$) than using $\approx 2M$ segments that degrades to 20.86 $\leftrightarrow$ 16.44 BLEU. However, for UNMT it is a common knowledge where more monolingual data leads to better performance (Lample and Conneau, 2019). We confirmed this by reducing the OOD to $200k$ as in ZNMT and SNMT, where we observed a 5 BLEU drop in UNMT performance in both $Gl \leftrightarrow En$ directions. For this reason, we train UNMT models using all the available IND and $\approx 2M$ OOD segments. In other words, the unsupervised models consume all the available IND and OOD monolingual data, that is $\times 10$ more than the SNMT baseline and our ZNMT utilized. In sum, this shows the efficiency of our approach to reach to a better performance with less resources. Detail comparisons are provided in the main experimental section.

Lastly, Fig. 5 shows an effective strategy of utilizing the IND and OOD data for ZNMT in a mixed domain (IOD) scenario. The test settings show first learning ZNMT with the IND and progressively incorporating OOD data (IND $>$ IOD) is the best approach, in comparison with (OOD $>$ IOD), or utilizing (IOD) from the beginning. Considering pre-trained models for ZNMT utilizes IND data (except for the unseen U), the finding is expected and leads to a better performance. Applying a similar (IND $>$ IOD) strategy for UNMT, however, resulted in a drop of up to 7 BLEU for $En \rightarrow Gl$, compared to training with the mixed domain (IOD) from the beginning. This is likely due to the fact that the the pre-training for UNMT observes both ID and OD data of $U-T$ ZSP and leading to a better learning when using IOD. In our main experimental setup we choose the best strategy for each of the approaches.

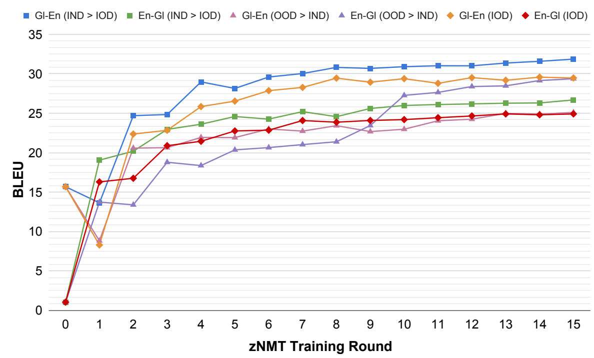

Figure 5: Training strategies to best utilize in-domain (IND) and out-of-domain (OOD) monolingual data.

C Model Configuration and Parameters

Model	Initialization	Params ($\times 10^6$)	Layers
MLM	-	41	6
BITM	-	53	6
MUTM*	-	69	6
NMT	-	38	4
SNMT	-	38	4
MNMT	-	69	6
UNMT	MLM	86	6
ZNMT	BITM	53	6
ZNMT	MUTM*	69	6

Table 5: Model, parameter size, and number of self-attention layers. MUTM* represents both MU100 and MU108.

To tackle over-fitting in the bilingual baseline supervised and semi-supervised NMT models we employ a dropout rate of 0.1 on the attention and 0.3 on all the other layers. Whereas the dropout rate for all the other models are set uniformly to 0.1. We use source and target tied embeddings (Press and Wolf, 2016). Samples exceeding 100 sub-word counts are discarded at time of training. Model training is done on a single V100 GPU with batch-size of $4,096$ tokens. Adam is used as an optimizer (Kingma and Ba, 2014) with a learning rate of 10^{-4}.

Details about model parameter are provided in Table 5. At time of training all models have shown to converge. While ZNMT shows the fastest learning curve within $15 - 20$ epochs, UNMT run up to 100 epochs to reach convergence.

D Languages and Data

Table 6 lists the languages and examples size from the TED talks data.

Language	Lang. Id	Train	Dev	Test
Arabic	ar	214111	4714	5953
Azerbaijani	az	5946	671	903
Belarusian	be	4509	248	664
Bulgarian	bg	174444	4082	5060
Bengali	bn	4649	896	216
Bosnian	bs	5664	474	463
Czech	cs	103093	3462	3831
Danish	da	44940	1694	1683
German	de	167888	4148	4491
Greek	el	134327	3344	4431
Esperanto	eo	6535	495	758
Spanish	es	196026	4231	5571
Estonian	et	10738	740	1087
Basque	eu	5182	318	379
Persian	fa	150965	3930	4490
Finnish	fi	24222	981	1301
French-Canadian	fr-ca	19870	838	1611
French	fr	192304	4320	4866
Galician	gl	10017	682	1007
Hebrew	he	211819	4515	5508
Hindi	hi	18798	854	1243
Croatian	hr	122091	3333	4881
Hungarian	hu	147219	3725	4981
Armenian	hy	21360	739	1567
Indonesian	id	87406	2677	3179
Italian	it	204503	4547	5625
Japanese	ja	204090	4429	5565
Georgian	ka	13193	654	943
Kazakh	kk	3317	938	775
Korean	ko	205640	4441	5637
Kurdish	ku	10371	265	766
Lithuanian	lt	41919	1791	1791
Macedonian	mk	25335	640	438
Mongolian	mn	7607	372	414
Marathi	mr	9840	767	1090
Malay	ms	5220	539	260
Burmese	my	21497	741	1504
Norwegian	nb	15825	826	806
Dutch	nl	183767	4459	5006
Polish	pl	176169	4108	5010
Portuguese-Brazilian	pt-br	184755	4035	4855
Portuguese	pt	51785	1193	1803
Romanian	ro	180484	3904	4631
Russian	ru	208458	4814	5483
Slovak	sk	61470	2271	2445
Slovenian	sl	19831	1068	1251
Albanian	sq	44525	1556	2443
Serbian	sr	136898	3798	4634
Swedish	sv	56647	1729	2283
Tamil	ta	6224	447	832
Thai	th	98064	2989	3713
Turkish	tr	182470	4045	5029
Ukrainian	uk	108495	3060	3751
Urdu	ur	5977	508	1006
Vietnamese	vi	171995	4645	4391
Chinese-China	zh-cn	199855	4558	5251
Chinese	zh	5534	547	494
Chinese-Taiwan	zh-tw	202646	4583	5377

Table 6: Languages and the parallel number of segments paired with English from the the TED Talks data (Qi et al., 2018). The four languages used as an unseen (U) source are highlighted.