Improved Zero-shot Neural Machine Translation via Ignoring Spurious Correlations

Jiatao Gu*[◇], Yong Wang*[†], Kyunghyun Cho[◇‡] and Victor O.K. Li[†]

[◇]Facebook AI Research
[†]The University of Hong Kong
[‡]New York University, CIFAR Azrieli Global Scholar
[◇]{jgu, kyunghyuncho}@fb.com
[†]{wangyong, vli}@eee.hku.hk

Abstract

Zero-shot translation, translating between language pairs on which a Neural Machine Translation (NMT) system has never been trained, is an emergent property when training the system in multilingual settings. However, naïve training for zero-shot NMT easily fails, and is sensitive to hyper-parameter setting. The performance typically lags far behind the more conventional pivot-based approach which translates twice using a third language as a pivot. In this work, we address the degeneracy problem due to *capturing spurious correlations* by quantitatively analyzing the mutual information between language IDs of the source and decoded sentences. Inspired by this analysis, we propose to use two simple but effective approaches: (1) decoder pre-training; (2) back-translation. These methods show significant improvement ($4 \sim 22$ BLEU points) over the vanilla zero-shot translation on three challenging multilingual datasets, and achieve similar or better results than the pivot-based approach.

1 Introduction

Despite the recent domination of neural network-based models (Sutskever et al., 2014; Bahdanau et al., 2014; Vaswani et al., 2017) in the field of machine translation, which have fewer pipelined components and significantly outperform phrase-based systems (Koehn et al., 2003), Neural Machine Translation (NMT) still works poorly when the available number of training examples is limited. Research on low-resource languages is drawing increasing attention, and it has been found promising to train a multilingual NMT (Firat et al., 2016a) model for high- and row-resource languages to deal with low-resource translation (Gu et al., 2018b). As an extreme in terms of the number of supervised examples, prior works dug into translation with zero-resource (Firat et al., 2016b; Chen et al., 2017; Lample et al., 2018a,b) where the language pairs in interest do not have any parallel corpora between them. In particular, Johnson et al. (2017) observed an emergent property of *zero-shot translation* where a trained multilingual NMT model is able to automatically do translation on unseen language pairs; we refer to this setting as zero-shot NMT from here on.

In this work, we start with a typical degeneracy issue of zero-shot NMT, reported in several recent works (Arivazhagan et al., 2018; Sestorain et al., 2018), that zero-shot NMT is sensitive to training conditions, and the translation quality usually lags behind the pivot-based methods which use a shared language as a bridge for translation (Utiyama and Isahara, 2007; Cheng et al., 2016; Chen et al., 2017). We first quantitatively show that this degeneracy issue of zero-shot NMT is a consequence of capturing spurious correlation in the data. Then, two approaches are proposed to help the model ignore such correlation: language model pre-training and back-translation. We extensively evaluate the effectiveness of the proposed strategies on four languages from Europarl, five languages from IWSLT and four languages from MultiUN. Our experiments demonstrate that the proposed approaches significantly improve the baseline zero-shot NMT performance and outperforms the pivot-based translation in some language pairs by $2 \sim 3$ BLEU points.

2 Background

Given a source sentence $x = \{x_1, ..., x_{T'}\}$, a neural machine translation model factorizes the distribution over output sentences $y = \{y_1, ..., y_T\}$ into a product of conditional probabilities:

$$p(y|x;\theta) = \prod_{t=1}^{T+1} p(y_t|y_{0:t-1}, x_{1:T'};\theta), \quad (1)$$

* Equal contribution.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1258–1268
Florence, Italy, July 28 - August 2, 2019.

where special tokens y_0 ($\langle\text{bos}\rangle$) and y_{T+1} ($\langle\text{eos}\rangle$) are used to represent the beginning and the end of a target sentence. The conditional probability is parameterized using a neural network, typically, an encoder-decoder architecture based on either RNNs (Sutskever et al., 2014; Cho et al., 2014; Bahdanau et al., 2014), CNNs (Gehring et al., 2017) or the Transformers (Vaswani et al., 2017).

Multilingual NMT We start with a many-to-many multilingual system similar to Johnson et al. (2017) which leverages the knowledge from translation between multiple languages. It has an identical model architecture as the single pair translation model, but translates between multiple languages. For a different notation, we use (x^i, y^j) where $i, j \in \{0, ..., K\}$ to represent a pair of sentences translating from a source language i to a target language j. $K+1$ languages are considered in total. A multilingual model is usually trained by maximizing the likelihood over training sets $D^{i,j}$ of all available language pairs $\mathcal{S}$. That is:

$$\max_{\theta} \frac{1}{|\mathcal{S}| \cdot |D^{i,j}|} \sum_{(x^i,y^j)\in D^{i,j},(i,j)\in\mathcal{S}} \mathcal{L}^j_\theta(x^i, y^j), \quad (2)$$

where we denote $\mathcal{L}^j_\theta(x^i, y^j) = \log p(y^j|x^i, \boldsymbol{j}; \theta)$. Specifically, the target language ID j is given to the model so that it knows to which language it translates, and this can be readily implemented by setting the initial token $y_0 = j$ for the target sentence to start with.[1] The multilingual NMT model shares a single representation space across multiple languages, which has been found to facilitate translating low-resource language pairs (Firat et al., 2016a; Lee et al., 2016; Gu et al., 2018b,c).

Pivot-based NMT In practise, it is almost impossible for the training set to contain all $K \times (K+1)$ combinations of translation pairs to learn a multilingual model. Often only one (e.g. English) or a few out of the $K+1$ languages have parallel sentence pairs with the remaining languages. For instance, we may only have parallel pairs between English & French, and Spanish & English, but not between French & Spanish. What happens if we evaluate on an unseen direction e.g. Spanish to French? A simple but commonly used solution is *pivoting*: we first translate from Spanish to English, and then from English to French with two separately trained single-pair models or a single multilingual model. However, it comes with two drawbacks: (1) at least $2\times$ higher latency than that of a comparable direct translation model; (2) the models used in pivot-based translation are not trained taking into account the new language pair, making it difficult, especially for the second model, to cope with errors created by the first model.

Zero-shot NMT Johnson et al. (2017) showed that a trained multilingual NMT system could automatically translate between unseen pairs without any direct supervision, as long as both source and target languages were included in training. In other words, a model trained for instance on English & French and Spanish & English is able to directly translate from Spanish to French. Such an emergent property of a multilingual system is called *zero-shot translation*. It is conjectured that zero-shot NMT is possible because the optimization encourages different languages to be encoded into a shared space so that the decoder is detached from the source languages. As an evidence, Arivazhagan et al. (2018) measured the "cosine distance" between the encoder's pooled outputs of each sentence pair, and found that the distance decreased during the multilingual training.

3 Degeneracy Issue of Zero-shot NMT

Despite the nice property of the emergent zero-shot NMT compared to other approaches such as pivot-based methods, prior works (Johnson et al., 2017; Firat et al., 2016b; Arivazhagan et al., 2018), however, have shown that the quality of zero-shot NMT significantly lags behind pivot-based translation. In this section, we investigate an underlying cause behind this particular degeneracy issue.

3.1 Zero-shot NMT is Sensitive to Training Conditions

Preliminary Experiments Before drawing any conclusions, we first experimented with a variety of hyper-parameters to train multilingual systems and evaluated them on zero-shot situations, which refer to language pairs without parallel resource.

We performed the preliminary experiments on Europarl[2] with the following languages: English (En), French (Fr), Spanish (Es) and German (De) with no parallel sentences between any two of Fr,

[1] Based on prior works (Arivazhagan et al., 2018), both options work similarly. Without loss of generality, we use the target language ID as the initial token y_0 of the decoder.

[2] http://www.statmt.org/europarl/

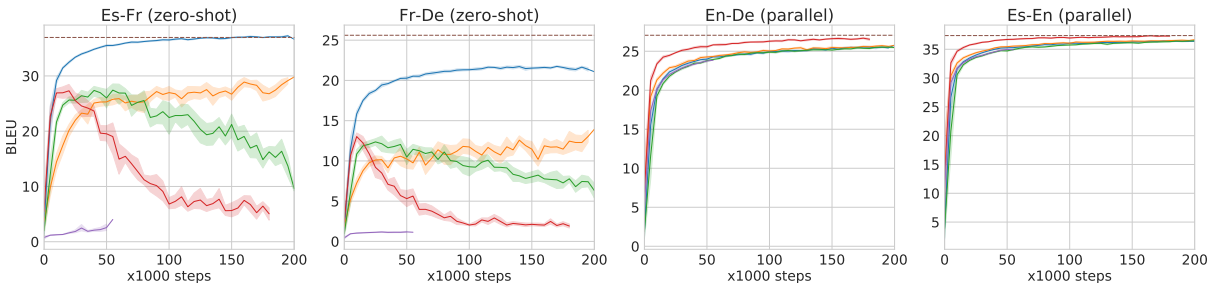


Figure 1: Partial results on zero-shot and parallel directions on Europarl dataset with variant multilingual training conditions (blue: `default`, red: `large-bs`, orange: `pytorch-init`, green: `attn-drop`, purple: `layerwise-attn`). The dashed lines are the pivot-based or direct translation results from baseline models.

Es and De. We used newstest2010[3] as the validation set which contains all six directions. The corpus was preprocessed with $40,000$ BPE operations across all the languages. We chose Transformer (Vaswani et al., 2017) – the state-of-the-art NMT architecture on a variety of languages – with the parameters of $d_{\text{model}} = 512$, $d_{\text{hidden}} = 2048$, $n_{\text{heads}} = 8$, $n_{\text{layers}} = 6$. Multiple copies of this network were trained on data with all parallel directions for {De,Es,Fr} & En, while we varied other hyper-parameters. As the baseline, six single-pair models were trained to produce the pivot results.

Results The partial results are shown in Fig. 1 including five out of many conditions on which we have tested. The `default` uses the exact Transformer architecture with *xavier_uniform* (Glorot and Bengio, 2010) initialization for all layers, and is trained with $\text{lr}_{\text{max}} = 0.005$, $t_{\text{warmup}} = 4000$, dropout $= 0.1$, $n_{\text{batch}} = 2400$ tokens/direction. For the other variants compared to the `default` setting, `large-bs` uses a bigger batch-size of 9,600; `attn-drop` has an additional dropout (0.1) on each attention head (Vaswani et al., 2017); we use the Pytorch's default method[4] to initialize all the weights for `pytorch-init`; we also try to change the conventional architecture with a layer-wise attention (Gu et al., 2018a) between the encoder and decoder, and it is denoted as `layerwise-attn`. All results are evaluated on the validation set using greedy decoding.

From Fig. 1, we can observe that the translation quality of zero-shot NMT is highly sensitive to the hyper-parameters (e.g. `layerwise-attn` completely fails on zero-shot pairs) while almost all the models achieve the same level as the baseline does on parallel directions. Also, even with the stable setting (`default`), the translation quality of zero-shot NMT is still far below that of pivot-based translation on some pairs such as Fr-De.

3.2 Performance Degeneracy is Due to Capturing Spurious Correlation

We look into this problem with some quantitative analysis by re-thinking the multilingual training in Eq. (4). For convenience, we model the decoder's output y^j as a combination of two factors: the output language ID $z \in \{0, \ldots, K\}$, and language-invariant semantics s (see Fig. 2 for a graphical illustration.). In this work, both z and s are unobserved variables before the y^j was generated. Note that z is not necessarily equal to the language id j.

The best practise for zero-shot NMT is to make z and s conditionally independent given the source sentence. That is to say, z is controlled by j and s is controlled by x^i. This allows us to change the target language by setting j to a desired language, and is equivalent to ignoring the correlation between x^i and z. That is, the mutual information between the source language ID i and the output language ID z – $I(i; z)$ – is **0**. However, the conventional multilingual training on an imbalanced dataset makes zero-shot NMT problematic because the MLE objective will try to capture all possible correlations in the data including the spurious dependency between i and z. For instance, consider training a multilingual NMT model for Es as input only with En as the target language. Although it is undesirable for the model to capture the dependency between i (Es) and z (En), MLE does not have a mechanism to prevent it (i.e., $I(i; z) > 0$) from happening. In other words, we cannot explicitly control the trade off between $I(i; z)$ and $I(j; z)$ with MLE training. When $I(i; z)$ increases as opposed to $I(j; z)$, the

[3] http://www.statmt.org/wmt18/translation-task.html

[4] We use https://pytorch.org/docs/master/_modules/torch/nn/modules/linear.html#Linear

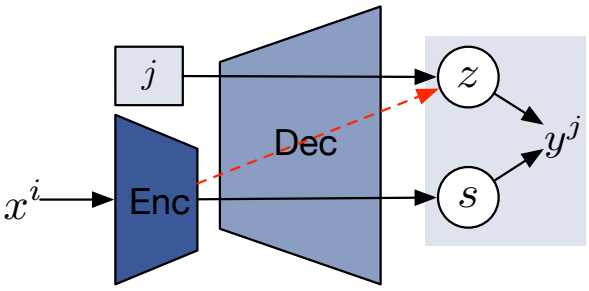


Figure 2: A conceptual illustration of decoupling the output translation (y^j) into two latent factors (language type and the semantics) where the undesired spurious correlation (in red) will be wrongly captured if i is always translated to j during training.

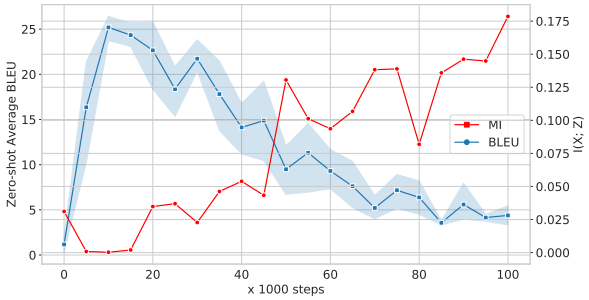


Figure 3: The learning curves of the mutual information between input and output language IDs as well as the averaged BLEU scores of all zero-shot directions on the validation sets for the `large-bs` setting.

decoder ignores j, which makes it impossible for the trained model to perform zero-shot NMT, as the decoder cannot output a translation in a language that was not trained before.

Quantitative Analysis We performed the quantitative analysis on the estimated mutual information $I(i;z)$ as well as the translation quality of zero-shot translation on the validation set. As an example, we show the results of `large-bs` setting in Fig. 3 where the $I(i;z)$ is estimated by:

$$I(i;z) \approx \frac{1}{(K+1)^2} \sum_{i,j} \log \left[\frac{\tilde{p}(z,i)}{\tilde{p}(z) \cdot \tilde{p}(i)} \right], \quad (3)$$

where the summation is over all possible language pairs, and $\tilde{p}(\cdot)$ represents frequency. The latent language identity $z = \phi(y^j)$ is estimated by an external language identification tool given the actual output (Lui and Baldwin, 2012). In Fig. 3, the trend of zero-shot performance is inversely proportional to $I(i;z)$, which indicates that the degeneracy is from the spurious correlation.

The analysis of the mutual information also explains the sensitivity issue of zero-shot NMT during training. As a side effect of learning translation, $I(i;z)$ tends to increase more when the training conditions make MT training easier (e.g. large batch-size). The performance of zero-shot NMT becomes more unstable and fails to produce translation in the desired language (j).

4 Approaches

In this section, we present two existing, however, not investigated in the scenario of zero-shot NMT approaches – decoder pre-training and back-translation – to address this degeneracy issue.

4.1 Language Model Pre-training

The first approach strengthens the decoder language model (LM) prior to MT training. Learning the decoder language model increases $I(j;z)$ which facilitates zero-shot translation. Once the model captures the correct dependency that guides the model to output the desired language, it is more likely for the model to ignore the spurious correlation during standard NMT training. That is, we pre-train the decoder as a multilingual language model. Similar to Eq. (4):

$$\max_{\theta} \frac{1}{|\mathcal{S}| \cdot |D^{i,j}|} \sum_{(x^i,y^j) \in D^{i,j}, (i,j) \in \mathcal{S}} \tilde{\mathcal{L}}^j_\theta(y^j), \quad (4)$$

where $\tilde{\mathcal{L}}^j_\theta(y^j) = \log p(y^j | \mathbf{0}, \boldsymbol{j}; \theta)$, which represents that pre-training can be implemented by simply replacing all the source representations by **zero** vectors during standard NMT training (Sennrich et al., 2016). In Transformer, it is equivalent to ignoring the attention modules between the encoder and decoder.

The proposed LM pre-training can be seen as a rough approximation of marginalizing all possible source sentences, while empirically we found it worked well. After a few gradient descent steps, the pre-trained model continues with MT training. In this work, we only consider using the same parallel data for pre-training. We summarize the pros and cons as follows:

Pros: Efficient (a few LM training steps + NMT training); no additional data needed;

Cons: The LM pre-training objective does not necessarily align with the NMT objective.

4.2 Back-Translation

In order to apply language model training along with the NMT objective, we have to take the encoder into account. We use back-translation (BT, Sennrich et al., 2016), but in particular for multilingual training. Unlike the original purpose of using BT for semi-supervised learning, we utilize BT to generate *synthetic* parallel sentences for all zero-shot directions (Firat et al., 2016b), and train the multilingual model from scratch on the merged datasets of both real and synthetic sentences. By doing so, every language is forced to translate to all the other languages. Thus, $I(i; z)$ is effectively close to 0 from the beginning, preventing the model from capturing the spurious correlation between i and z.

Generating the synthetic corpus requires at least a reasonable starting point that translates on zero-shot pairs which can be chosen either through a pivot language (denoted as BTTP) or the current zero-shot NMT trained without BT (denoted BTZS). For instance, in previous examples, to generate synthetic pairs for Es-Fr given the training set of En-Fr, BTTP translates every En sentence to Es with a pre-trained En-Es model (used in pivot-based MT), while BTZS uses the pre-trained zero-shot NMT to directly translate all Fr sentences to Es. Next, we pair the generated sentences in the reverse direction Es-Fr and merge them to the training set. The same multilingual training is applied after creating synthetic corpus for all translation pairs. Similar methods have also been explored by Firat et al. (2016b); Zheng et al. (2017); Sestorain et al. (2018), but have not been studied or used in the context of zero-shot NMT.

Pros: BT explicitly avoids the spurious correlation. Also, BTZS potentially improves further by utilizing the zero-shot NMT model augmented with LM pre-training.

Cons: BT is computationally more expensive as we need to create synthetic parallel corpora for all language pairs (up to $O(K^2)$) to train a multilingual model for K languages; both the performance of BTTP and BTZS might be affected by the quality of the pre-trained models.

5 Experiments

5.1 Experimental Settings

Dataset We extensively evaluate the proposed approaches (LM, BTTP, BTZS) on three multilingual datasets across a variety of languages: Europarl, IWSLT[5] and MultiUN.[6] The detailed statistics of the training set are in Table 1, where we simulate the zero-shot settings by only allowing parallel sentences from/to English. With IWSLT, we also simulate the scenario of having a chain of pivot languages (IWSLT-b). Also, another additional dataset (Europarl-b) is included where the zero-shot pairs have neither direct nor pivot parallel sentences (similar to unsupervised translation). In such cases, we expect pivot-based methods (including the proposed BTTP) are not applicable. We use the standard validation and test sets to report the zero-shot performance. Besides, we preprocess all the datasets following the protocol used in the preliminary experiments.

Dataset	parallel pairs	size/pair
Europarl	Es-En, De-En, Fr-En	2M
Europarl-b	Es-En, Fr-De	1.8M
IWSLT	De-En, It-En, Nl-En, Ro-En	.22M
IWSLT-b	De-En, En-It, It-Ro, Ro-Nl	.22M
MultiUN	Ar-En, Ru-En, Zh-En	2M

Table 1: Overall dataset statistics where each pair has a similar number of examples shown in the rightmost column (we sub-sampled $2M$ sentences per language pair for MultiUN). All the remaining directions are used to evaluate the performance of zero-shot NMT.

Training Conditions For all non-IWSLT experiments, we use the same architecture as the preliminary experiments with the training conditions of `default`, which is the most stable setting for zero-shot NMT in Sec. 3.1. Since IWSLT is much smaller compared to the other two datasets, we find that the same hyper-parameters except with $t_{\text{warmup}} = 8000$, dropout $= 0.2$ works better.

Models As the baseline, two pivot-based translation are considered:

- PIV-S (through two single-pair NMT models trained on each pair;)
- PIV-M (through a single multilingual NMT model trained on all available directions;)

Moreover, we directly use the multilingual system that produce PIV-M results for the vanilla zero-shot NMT baseline.

[5] https://sites.google.com/site/iwsltevaluation2017
[6] http://opus.nlpl.eu/MultiUN.php

As described in Sec. 4, both the LM pre-training and BT use the same datasets as that in MT training. By default, we take the checkpoint of $20,000$ steps LM pre-training to initialize the NMT model as our preliminary exploration implied that further increasing the pre-training steps would not be helpful for zero-shot NMT. For BTTP, we choose either PIV-S or PIV-M to generate the synthetic corpus based on the average BLEU scores on parallel data. On the other hand, we always select the best zero-shot model with LM pre-training for BTZS by assuming that pre-training consistently improves the translation quality of zero-shot NMT.

5.2 Model Selection for Zero-shot NMT

In principle, zero-shot translation assumes we cannot access any parallel resource for the zero-shot pairs during training, including cross-validation for selecting the best model. However, according to Fig. 1, the performance of zero-shot NMT tends to drop while the parallel directions are still improving which indicates that simply selecting the best model based on the validation set of parallel directions is sub-optimal for zero-shot pairs. In this work, we propose to select the best model by maximizing the likelihood over all available validation set $\hat{D}^{i,j}$ of parallel directions together with a language model score from a fully trained language model θ' (Eq. (4)). That is,

$$\sum_{\substack{(x^i,y^j)\in\hat{D}^{i,j}\\(i,j)\in\mathcal{S}}}\left[\mathcal{L}^j_\theta(x^i,y^j)+\sum_{\substack{(i,k)\notin\mathcal{S}\\i\neq k}}\frac{\tilde{\mathcal{L}}^k_{\theta'}(\hat{y}^k)}{K-|\mathcal{S}|}\right], \quad (5)$$

where $\hat{y}^k$ is the greedy decoding output generated from the current model $p(\cdot|x^i, \boldsymbol{k}; \theta)$ by forcing it to translate x^i to language k that has no parallel data with i during training. The first term measures the learning progress of machine translation, and the second term shows the level of degeneracy in zero-shot NMT. Therefore, when the spurious correlation between the input and decoded languages is wrongly captured by the model, the desired language model score will decrease accordingly.

5.3 Results and Analysis

Overall Performance Comparison We show the translation quality of zero-shot NMT on the three datasets in Table 2. All the results (including pivot-based approaches) are generated using beam-search with beam size $= 4$ and length penalty $\alpha = 0.6$ (Vaswani et al., 2017). Experimental results in Table 2 demonstrate that both our proposed approaches achieve significant improvement in zero-shot translation for both directions in all the language pairs. Only with LM pre-training, the zero-shot NMT has already closed the gap between the performance and that of the strong pivot-based baseline for datasets. For pairs which are lexically more similar compared to the pivot language (e.g. Es-Fr v.s. En), ZS+LM achieved much better performance than its pivot-based counterpart. Depending on which languages we consider, zero-shot NMT with the help of BTTP & BTZS can achieve a significant improvement around $4 \sim 22$ BLEU points compared to the naïve approach. For a fair comparison, we also re-implement the alignment method proposed by Arivazhagan et al. (2018) based on cosine distance and the results are shown as ZS+Align in Table. 2, which is on average 1.5 BLEU points lower than our proposed ZS+LM approach indicating that our approaches might fix the degeneracy issue better.

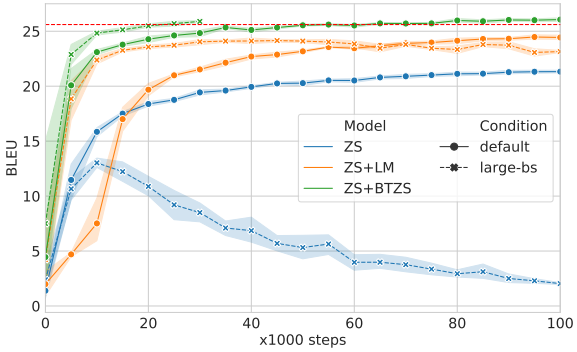


Figure 4: Learning curves of the two proposed approaches (LM, BTZS) and the vanilla ZS on Europarl Fr→De with two conditions (`default`, `large-bs`). The red dashed line is the pivot-based baseline.

As a reference of upper bound, we also include the results with a fully supervised setting, where all the language pairs are provided for training. Table 2 shows that the proposed BTTP & BTZS are competitive and even very close to this upper bound, and BTZS is often slightly better than BTTP across different languages.

No Pivots We conduct experiments on the setting without available pivot languages. Shown in Table 2(b), our training sets only contain Es-En and De-Fr. Then if we want to translate from De to Fr, pivot-based methods will not work. However, we can still perform zero-shot NMT by simply training a multilingual model on the

Europarl	(a) De, Es, Fr ↔ En								(b) Es ↔ En, Fr ↔ De			
Model	De-Es		De-Fr		Es-Fr		Zero	Parallel	Es-Fr		De-En	
	←	→	←	→	←	→	Avg	Avg	←	→	←	→
PIV-S	26.2	31.2	**25.9**	32.2	35.7	38.0	31.5	**35.0**	– not applicable –			
PIV-M	26.2	31.1	25.2	31.5	35.4	37.1	31.1	34.4	– not applicable –			
ZS	22.1	30.2	21.7	29.6	36.2	36.7	29.4	34.4	29.5	27.5	14.3	23.7
ZS+Align (2018)	24.7	31.4	23.8	31.0	37.3	38.5	31.1	34.5	–	–	–	–
ZS+LM	25.9	32.8	25.5	32.3	**39.3**	**40.0**	32.6	34.6	**34.9**	**37.1**	**21.5**	**30.0**
ZS+BTTP	**27.1**	**33.0**	**26.4**	**33.0**	39.1	**40.0**	**33.1**	33.9	– not applicable –			
ZS+BTZS	**26.7**	**33.2**	**25.9**	**33.1**	**40.0**	**41.4**	**33.4**	**34.7**	**39.7**	**40.5**	**25.1**	**30.6**
Full	28.5	34.1	27.9	34.2	40.0	42.0	34.4	34.8	40.0	42.0	27.0	33.4

IWSLT	(c) De, It, Nl, Ro ↔ En													
Model	De-It		De-Nl		De-Ro		It-Nl		It-Ro		Nl-Ro		Zero	Parallel
	←	→	←	→	←	→	←	→	←	→	←	→	Avg	Avg
PIV-S	16.7	16.3	19.1	17.7	17.5	15.0	18.4	18.6	18.8	17.2	18.3	17.0	17.6	29.8
PIV-M	21.4	21.6	24.0	23.7	22.3	20.0	22.7	22.4	23.6	21.3	23.0	21.1	22.3	35.0
ZS	14.8	17.2	16.7	17.8	14.9	16.6	18.4	16.1	19.7	17.8	16.2	17.5	17.0	35.0
ZS+LM	21.3	20.9	24.7	24.1	22.3	19.8	22.2	22.3	23.2	22.1	23.0	21.6	22.3	34.9
ZS+BTTP	**23.3**	**23.3**	**26.5**	**25.8**	**23.9**	**22.1**	**24.6**	**24.3**	**25.9**	**23.7**	**24.7**	**23.7**	**24.3**	**35.2**
ZS+BTZS	**22.6**	**23.3**	**27.2**	**26.5**	**23.6**	**21.8**	**24.3**	**24.0**	**25.7**	**23.6**	**25.4**	**23.3**	**24.3**	**35.5**
Full	23.9	23.9	27.0	26.1	24.8	22.7	25.6	24.6	25.9	24.2	25.1	23.9	24.8	35.7

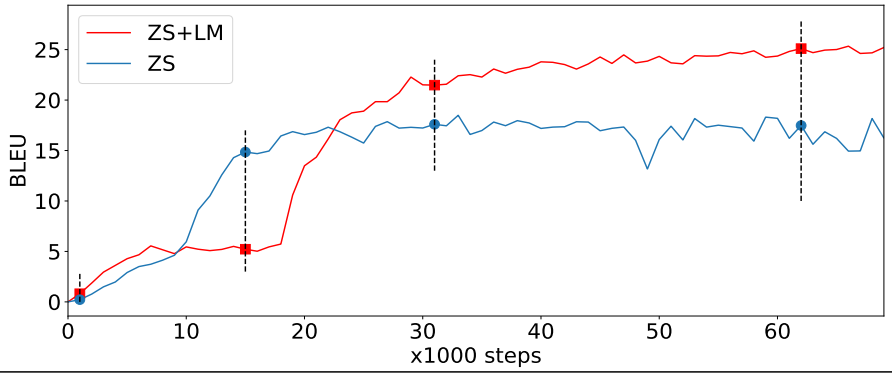
IWSLT	(d) De ↔ En ↔ It ↔ Ro ↔ Nl				MultiUN	(e) Ar, Ru, Zh ↔ En							
Model	De-It		De-Nl		Model	Ar-Ru		Ar-Zh		Ru-Zh		Zero	Parallel
	←	→	←	→		←	→	←	→	←	→	Avg	Avg
PIV-S	16.7	16.3	–	–	PIV-S	**31.4**	**33.5**	**31.2**	**50.4**	**31.2**	**48.0**	**37.6**	**48.4**
PIV-M	**22.7**	22.0	18.8	18.3	PIV-M	28.4	29.9	27.7	45.7	27.2	44.2	33.8	44.5
ZS	21.3	21.0	23.9	24.0	ZS	15.6	12.7	16.7	17.0	12.8	14.9	15.0	44.5
ZS+LM	22.2	**22.2**	**25.0**	**24.6**	ZS+LM	28.0	21.5	27.3	43.8	19.9	43.3	30.6	45.8
ZS+BTTP	–	–	–	–	ZS+BTTP	**31.0**	31.7	30.1	48.2	29.9	46.4	36.2	45.7
ZS+BTZS	**22.9**	**22.9**	**26.8**	**26.2**	ZS+BTZS	**31.4**	**33.1**	**31.1**	**49.4**	**30.8**	**46.8**	**37.1**	**47.4**
Full	23.9	23.9	27.0	26.1	Full	31.7	32.5	30.8	49.1	29.5	47.2	36.8	45.6

Table 2: Overall BLEU scores including parallel and zero-shot directions on the test sets of three multilingual datasets. In (a) (c) (e), En is used as the pivot-language; no language is available as the pivot for (b); we also present partial results in (d) where a chain of pivot languages are used. For all columns, the highest **two** scores are marked in bold for all models except for the fully-supervised "upper bound".

merged dataset. As shown in Table 2(a) and (b), although the setting of no pivot pairs performs slightly worse than that with pivot languages, both our approaches (LM, BTZS) substantially improve the vanilla model and achieve competitive performance compared to the fully supervised setting.

A Chain of Pivots We analyze the case where two languages are connected by a chain of pivot languages. As shown in Table 1(IWSLT-b), we used IWSLT which contains pairs for De-En, En-It, It-Ro, Ro-Nl. If we translate from De to Nl with pivot-based translation, pivoting from a chain of languages (De-En-It-Ro-Nl) is required, which suffers from computational inefficiency and error accumulation. In such cases, however, zero-shot NMT is able to directly translate between any two languages. Table 2(d) shows that the performance of pivot-based methods dramatically degrades as the length of the chain increases, while ZS does not have this degradation and still achieves large gains compared to the pivot-based translation.

Robustness Analysis From Fig. 4, we show the learning curves of zero-shot NMT with and without our proposed methods. Both the models with LM pre-training and BTZS show robustness in two conditions and achieve competitive and even better results than the pivot-based translation, while the vanilla model is unstable and completely fails



	SOURCE	менее значительные изменения в инде@@ кс@@ е развития человеческого потенциала еще больше сни@@ з@@ или довер@@ ие к нему и веду@@ щую роль , которую могли играть доклады о развитии человеческого потенциала в оценке уровня развития человека .
	TARGET	人类 发展 指数 的 小@@ 幅 修订 进一步 损害 了 人类 发展 报告 对 衡量 人类 发展 的 信@@ 誉@@ 度 和 领导@@ 力 。
step 1000	ZS	在 在 在 在 在 在 在 在 在 在 在 在 在 的 的 的 的 的 ， 在 在 在 在 在 在 在 在 在 在 在 在 在 在 在 在 的 的 的 的 的 的 。
	ZS+LM	在 这方面 ， 我们 强调 了 在 发展 筹资 问题 国际 会议 上 取得 的 进展 ， 并 强调 了 在 发展 筹资 问题 国际 会议 上 取得 的 进展 ， 包括 在 发展 筹资 问题 上 取得 了 进展 。
step 15000	ZS	在 人类 发展 指数 中 ， 更 重要 的 发展 也 更加 严重 ， 也 更加 脆弱 ， 也 是 在 评估 人类 发展 的 情况 下 ， 可 发挥 重要 作用 。
	ZS+LM	在 发展 中 的 人类 发展 方面 ， 可以 发挥 更大 的 作用 ， 并 使 其 能够 在 其 报告 中 发挥 更大 的 作用 ， 并 使 其 能够 在 发展 中 的 能力 建设 。
step 31000	ZS	人类 发展 index 的 重大 变化 也 进一步 降低 了 对 人类 发展 的 能力 ， 并 发挥 了 重要 作用 ， 可 在 人类 发展 评估 方面 提供 human development 报告 。
	ZS+LM	人类 发展 指数 的 相对 较 小 的 变化 进一步 减少 了 它 的 信心 ， 并 使 它 能够 发挥 人类 发展 报告 的 主@@ 导@@ 作用 。
step 62000	ZS	less significant changes in the human development index have further reduced its credibility and leadership role in human development assessment reports .
	ZS+LM	人类 发展 指数 的 较 小 变化 进一步 减少 了 人们 对 人类 发展 的 信任 ， 并 发挥 了 对 人类 发展 水平 的 评价 作用 。

Figure 5: Zero-shot translation performance on Ru → Zh from MultiUN dataset. (↑) An example randomly selected from the validation set, is translated by both the vanilla zero-shot NMT and that with LM pre-training at four checkpoints. Translation in an incorrect language (English) is marked in pink color. (←) We showed the two learning curves for the averaged zero-shot BLEU scores on validation set of Multi-UN with the corresponded checkpoints marked.

after a small number of iterations on `large-bs`.

Case Study We also show a randomly selected example for Ru → Zh from the validation set of MultiUN dataset in Fig. 5. We can see that at the beginning, the output sentence of ZS+LM is fluent while ZS learns translation faster than ZS+LM. Then, En tokens starts to appear in the output sentence of ZS, and it totally shifts to En eventually.

6 Related Works

Zero-shot Neural Machine Translation Zero-shot NMT has received increasingly more interest in recent years. Platanios et al. (2018) introduced the contextual parameter generator, which generated the parameters of the system and performed zero-shot translation. Arivazhagan et al. (2018) conjectured the solution towards the degeneracy in zero-shot NMT was to guide an NMT encoder to learn language agnostic representations. Sestorain et al. (2018) combined dual learning to improve zero-shot NMT. However, unlike our work, none of these prior works performed quantitative investigation of the underlying cause.

Zero Resource Translation This work is also closely related to *zero-resource translation* which is a general task to translate between languages without parallel resources. Possible solutions include *pivot-based* translation, *multilingual* or *unsupervised* NMT. For instance, there have been attempts to train a single-pair model with a pivot-language (Cheng et al., 2016; Chen et al., 2017) or a pivot-image (Lee et al., 2017; Chen et al., 2018).

Unsupervised Translation Unlike the focus of this work, unsupervised translation usually refers to a zero-resource problem where many monolingual corpora are available. Lample et al. (2018a); Artetxe et al. (2018) proposed to enforce a shared

latent space to improve unsupervised translation quality which was shown not necessary by Lample et al. (2018b) in which a more effective initialization method for related languages was proposed.

Neural Machine Translation Pre-training As a standard transfer learning approach, pre-training significantly improves the translation quality of low resource languages by fine-tuning the parameters trained on high-resource languages (Zoph et al., 2016; Gu et al., 2018c; Lample and Conneau, 2019). Our proposed LM pre-training can also be included in the same scope while following a different motivation.

7 Conclusion

In this paper, we analyzed the issue of zero-shot translation quantitatively and successfully close the gap of the performance of between zero-shot translation and pivot-based zero-resource translation. We proposed two simple and effective strategies for zero-shot translation. Experiments on the Europarl, IWSLT and MultiUN corpora show that our proposed methods significantly improve the vanilla zero-shot NMT and consistently outperform the pivot-based methods.

Acknowledgement

This research was supported in part by the Facebook Low Resource Neural Machine Translation Award. This work was also partly supported by Samsung Advanced Institute of Technology (Next Generation Deep Learning: from pattern recognition to AI) and Samsung Electronics (Improving Deep Learning using Latent Structure). KC thanks support by eBay, TenCent, NVIDIA and CIFAR.

References

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2018. The missing ingredient in zero-shot neural machine translation.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *Proceedings of International Conference on Learning Representations (ICLR)*, Vancouver, Canada.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Yun Chen, Yang Liu, Yong Cheng, and Victor OK Li. 2017. A teacher-student framework for zero-resource neural machine translation. *arXiv preprint arXiv:1705.00753*.

Yun Chen, Yang Liu, and Victor OK Li. 2018. Zero-resource neural machine translation with multi-agent communication game. *arXiv preprint arXiv:1802.03116*.

Yong Cheng, Yang Liu, Qian Yang, Maosong Sun, and Wei Xu. 2016. Neural machine translation with pivot languages. *arXiv preprint arXiv:1611.04928*.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–Decoder approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016a. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *NAACL*.

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T Yarman Vural, and Kyunghyun Cho. 2016b. Zero-resource translation with multi-lingual neural machine translation. In *EMNLP*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of International Conference on Machine Learning (ICML)*.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. 2018a. Non-autoregressive neural machine translation. *ICLR*.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. 2018b. Universal neural machine translation for extremely low resource languages. *arXiv preprint arXiv:1802.05368*.

Jiatao Gu, Yong Wang, Yun Chen, Kyunghyun Cho, and Victor OK Li. 2018c. Meta-learning for low-resource neural machine translation. *arXiv preprint arXiv:1808.08437*.

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.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *Proceedings of International Conference on Learning Representations (ICLR)*, Vancouver, Canada.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049. Association for Computational Linguistics.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2016. Fully character-level neural machine translation without explicit segmentation. *arXiv preprint arXiv:1610.03017*.

Jason Lee, Kyunghyun Cho, Jason Weston, and Douwe Kiela. 2017. Emergent translation in multi-agent communication. *arXiv preprint arXiv:1710.06922*.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30. Association for Computational Linguistics.

Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018. Contextual parameter generation for universal neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 425–435. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for wmt 16. *arXiv preprint arXiv:1606.02891*.

Lierni Sestorain, Massimiliano Ciaramita, Christian Buck, and Thomas Hofmann. 2018. Zero-shot dual machine translation. *arXiv preprint arXiv:1805.10338*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *NIPS*.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 484–491.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*.

Hao Zheng, Yong Cheng, and Yang Liu. 2017. Maximum expected likelihood estimation for zero-resource neural machine translation. *IJCAI*.

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. Association for Computational Linguistics.

A Additional Experiments

A.1 Trade-off between decoding speed and translation quality

In Table. 3, we empirically tested the decoding speed by using either pivot-based methods or zero-shot NMT. The overhead of switching models in pivot-based translation has been ignored. All the speed are measured as "ms/sentence" and tested in parallel on 8 V100 GPUs using beam-search with a beam size 4.

Model	BLEU	Speed
PIV-S (greedy)	31.1	8.3
PIV-M (greedy)	30.6	8.3
PIV-S	31.5	13.3
PIV-M	31.1	13.3
ZS	29.4	**6.6**
ZS+LM	32.6	**6.6**
ZS+BTTP	33.1	**6.6**
ZS+BTZS	**33.4**	**6.6**

Table 3: Decoding speed and the translation quality (average BLEU scores) of the zero-shot pairs on Europarl dataset.

Vanilla zero-shot NMT is faster but performs worse than pivot-based methods. There exists a trade-off between the decoding speed and the translation quality where we also present a fast pivoting method where we found that using greedy-decoding for the pivot language only affects the translation quality by a small margin.

However, both our proposed approaches significantly improve the zero-shot NMT and outperforms the pivot-based translation with shorter decoding time, making such trade-off meaningless.

A.2 Effect of Using Multi-way Data

Prior research (Cheng et al., 2016) also reported that the original Europarl dataset contains a large proportion of multi-way translations. To investigate the affects, we followed the same process in (Cheng et al., 2016; Chen et al., 2017) to exclude all multi-way translation sentences, which means there are no overlaps in pairwise language pairs. The statistics of this modified dataset (Europarl-c) compared to the original Europarl dataset are shown in Table 4. Although we observed a performance drop by using data without multi-way sentences, the results in Table 5 show that the proposed LM pre-training is not affected by obtaining multi-way data and consistently improves the vanilla zero-shot NMT. We conjecture that the performance drop is mainly because of the size of the dataset. Also our methods can easily beat (Chen et al., 2017) with large margins.

Dataset	parallel pairs	size/pair
Europarl	Es-En, De-En, Fr-En	2M
Europarl-c	Es-En, De-En, Fr-En	.8M

Table 4: Europarl denotes multi-way dataset; Europarl-c denotes non multi-way dataset.

Model	Es→Fr		De→Fr	
	Yes	No	Yes	No
PIV-S	37.95	32.98	32.20	27.94
PIV-M	37.15	35.08	31.46	29.78
ZS	36.69	33.22	29.59	26.91
ZS + LM	**40.04**	**37.22**	**33.24**	**30.45**
Chen et al. (2017)	–	33.86	–	27.03

Table 5: Effects of multi-way data on Europarl. "Yes" means with multi-way translation, and "No" means the opposite.