Rethinking Zero-shot Neural Machine Translation: From a Perspective of Latent Variables

Weizhi Wang[1], **Zhirui Zhang**[2]*, **Yichao Du**[3], **Boxing Chen**[2], **Jun Xie**[2], **and Weihua Luo**[2]
[1]Rutgers University, New Brunswick, USA
[2]Machine Intelligence Technology Lab, Alibaba DAMO Academy
[3]University of Science and Technology of China, China
[1]weizhi.wang@rutgers.edu [2]zrustc11@gmail.com
[2]{boxing.cbx, qingjing.xj, weihua.luowh}@alibaba-inc.com
[3]duyichao@mail.ustc.edu.cn

Abstract

Zero-shot translation, directly translating between language pairs unseen in training, is a promising capability of multilingual neural machine translation (NMT). However, it usually suffers from capturing spurious correlations between the output language and language invariant semantics due to the maximum likelihood training objective, leading to poor transfer performance on zero-shot translation. In this paper, we propose to introduce a denoising autoencoder objective based on pivot language into traditional training objective to improve the translation accuary on zero-shot directions. The theoretical analysis from the perspective of latent variables shows that our proposed approach actually implicitly maximizes the probability distributions for zero-shot translation direction. On two benchmark machine translation datasets, we demonstrate that the proposed method is able to effectively eliminate the spurious correlations and significantly outperforms state-of-the-art methods with a remarkable performance.

1 Introduction

Multilingual neural machine translation (NMT) system concatenates multiple language pairs into one single neural-based model, enabling translation on multiple language directions (Firat et al., 2016; Ha et al., 2016; Johnson et al., 2017; Kudugunta et al., 2019; Arivazhagan et al., 2019b; Zhang et al., 2020). Besides, the multilingual NMT system can achieve translation on unseen language pairs in training, and we refer to this setting as zero-shot NMT. This finding is promising that zero-shot translation halves the decoding time of pivot-based method and avoids the problem of error propagation. Meanwhile, zero-shot NMT casts off the requirement of parallel data for a potentially quadratic number of language pairs, which is sometimes impractical especially between low-resource languages. Despite the potential benefits, achieving high-quality zero-shot translation is a very challenging task. Standard multilingual NMT systems are sensitive to hyper-parameter setting and tend to generate poor outputs.

Model	BLEU on DE⇒FR
DE⇒EN+EN⇒FR	6.0
PIV-(DE⇒EN+EN⇒FR)	31.5

Table 1: BLEU scores [%] of training multilingual NMT with these two translation directions and its pivoting variant on Europarl Dataset.

One line of research believes that the success of zero-shot translation depends on the ability of the model to learn language invariant features, or an interlingua, for cross-lingual transfer (Arivazhagan et al., 2019a; Ji et al., 2020; Liu et al., 2020). Arivazhagan et al. (2019a) design auxiliary losses on the NMT encoder that impose representational invariance across languages. Ji et al. (2020) build up a universal encoder for different languages via bridge language model pre-training, while Liu et al. (2020) disentangle positional information in multilingual NMT to obtain language-agnostic representations. On the other hand, Gu et al. (2019) point out that the conventional multilingual NMT model heavily captures spurious correlations between the output language and language invariant semantics due to the maximum likelihood training objective, making it hard to generate a reasonable translation in an unseen language. Then they investigate the effectiveness of decoder pre-training and back-translation on this issue.

In this paper, we focus on English-centric multilingual NMT and propose to incorporate a simple denoising autoencoder objective based on English language into traditional training objective of multilingual NMT to achieve better translation on zero-shot directions. This approach is motivated by an

[*] Corresponding author.

observation that: as shown in Table 1, if we only optimize two translation directions DE⇒EN and EN⇒FR in single model, it could not achieve zero-shot translation on DE⇒FR. It is because that the model easily learns high mutual information between language semantic of German and output language, ignoring the functionality of language IDs. Actually, this mutual information can be significantly reduced by directly replacing the original German sentence with a noisy target English sentence in training data, thereby guiding the model to learn the correct mapping between language IDs and output language. Also, we further analyze our proposed method by treating pivot language as latent variables and find that our proposed approach actually implicitly maximizes the probability distributions for zero-shot translation direction.

We verify the proposed method on two public multilingual datasets with several English-centric language-pairs, Europarl (Koehn, 2005) and MultiUN (Ziemski et al., 2016). Experimental results show that our proposed method not only achieves significant improvement over vanilla multilingual NMT on zero-shot directions, but also outperforms previous state-of-the-art methods.

2 Multilingual NMT

The multilingual NMT system (Johnson et al., 2017) combines different language directions into one single translation model. Due to data limitations of non-English languages, multilingual NMT system are mostly trained on large-scale English-centric corpus via maximizing the likelihood over all available language pairs $\mathcal{S}$:

$$\mathcal{L}_m(\theta) = \sum_{(i,j)\in\mathcal{S},(x,y)\in D^{i,j}} \log P(y|x, \mathbf{j}; \theta), \quad (1)$$

where $(i, j) \in \mathcal{S}$ are the sampled source language ID and target language ID in all available language pairs, $D^{i,j}$ represents for the corresponding parallel data, and θ is the model parameters. The target language ID is added as the initial token of source sentences, to let the model know which language it should translate to. Besides, multilingual NMT system has proven the capability of translating on unseen pairs in training (Firat et al., 2016; Johnson et al., 2017), which is property of **zero-shot translation**. However, the zero-shot translation quality significantly falls behind that of pivoting methods. The main issue leading to the unsatisfactory performance is that the multilingual NMT model captures spurious correlations between the output language and language invariant semantics due to the maximum likelihood training objective (Gu et al., 2019).

3 Method

In this section, we first introduce the denoising autoencoder task and then analyze the effectiveness of our proposed method from the perspective of latent variables.

Denoising Autoencoder Task. Given English-centric parallel data (X/Y/...⇔EN), we usually optimize the maximum likelihood training objective to build multilingual NMT model. Since the target language ID is inserted at the beginning of the source sentence and only treated as single token, the maximum likelihood training objective easily ignores the the functionality of target language ID, leading to unreasonable mutual information between language semantic of "X/Y/..." and output language of English. To this end, we introduce denoising sequence-to-sequence task, in which we directly replace the original input sentence with a noisy target English sentence in training data. In this way, previous mutual information can be significantly reduced, while enhancing the relationship between language IDs and output language. Specifically, we follow BART task (Lewis et al., 2020) and simply use all English sentences in parallel data to construct denoising English corpus D_{EN} via **Text Infilling**. Then we optimize the multilingual NMT model via maximizing the original translation objective $\mathcal{L}_m(\theta)$ and denoising autoencoder objective $\mathcal{L}_d(\theta)$:

$$\mathcal{L}_d(\theta) = \sum_{j=\text{<2en>},(\overline{y},y)\in D_{\text{EN}}} \log P(y|\overline{y}, \mathbf{j}; \theta) \quad (2)$$

$$\mathcal{L}_a(\theta) = \mathcal{L}_m(\theta) + \mathcal{L}_d(\theta). \quad (3)$$

Latent Variable Perspective. As for zero-shot translation, we actually aim at directly modeling probability distribution between non-English languages "X/Y/..." in the multilingual NMT system. For convenience, we consider the probability distribution $P(Y|X; D^*)$ between two non-English languages over the ideal parallel training data D^*. However, it is difficult for us to obtain such training data D^* for model training in practice. To address this issue, we convert maximizing $P(Y|X; D^*)$ into optimizing three existing sub-tasks by treating the English language as a latent variable h and introducing the probability distribution $P(h|\overline{h})$ of denoising autoencoder task:

$$\begin{aligned}
&P(Y|X;D^*) = \sum_{(x,y)\in D^*} \log P(y|x)\\
&= \sum_{(x,y)\in D^*} \log \sum_{h} P(y|h,x)P(h|x)\\
&\approx \sum_{(x,y)\in D^*} \log \sum_{h} P(h|\bar{h}) \frac{P(y|h)P(h|x)}{P(h|\bar{h})}\\
&\geq \sum_{(x,y)\in D^*} \sum_{h} P(h|\bar{h}) \log \frac{P(y|h)P(h|x)}{P(h|\bar{h})}\\
&= \sum_{(x,y)\in D^*} \mathbb{E}_{h\sim P(h|\bar{h})} \log P(y|h)\\
&\qquad\qquad - \mathbb{KL}(P(h|\bar{h})||P(h|x))\\
&= P^*(Y|X;D^*,P(h|\bar{h})),
\end{aligned} \tag{4}$$

where assuming $P(y|h,x) \approx P(y|h)$ due to the semantic equivalence of languages h and x. With above equation, the original objective is transformed into optimizing three sub-tasks $P(h|x)$, $P(y|h)$ and $P(h|\bar{h})$. Thus, incorporating the denoising autoencoder objective into the translation objective of multilingual NMT model help minimize the KL-divergence terms, which implicitly maximizes the probability distributions for zero-shot translation direction. Besides, following Ren et al. (2018), the gap between $P^*(Y|X;D^*)$ and $P(Y|X;D^*)$ is calculated as follow:

$$\begin{aligned}
&P(Y|X;D^*) - P^*(Y|X;D^*,P(h|\bar{h}))\\
&= \sum_{(x,y)\in D^*} \sum_{h} P(h|\bar{h}) \log \frac{P(h|\bar{h})P(y|x)}{P(y|h)P(h|x)}\\
&\approx \sum_{(x,y)\in D^*} \mathbb{KL}(P(h|\bar{h})||P(h|y)).
\end{aligned} \tag{5}$$

Once we complement $P(h|y)$ into three sub-tasks mentioned before, this gap could be further reduced, resulting in better performance on zero-shot translation directions.

4 Experiments

4.1 Experimental Settings

Datasets. We verify the proposed method on two benchmark machine translation datasets, Europarl and MultiUN. BLEU (Papineni et al., 2002) is used as the only metric for evaluating translation quality. For Europarl dataset, we select three European languages, Germany (De), French (Fr) and English (En). We remove all parallel sentences between De and Fr to ensure the zero-shot setting. We use WMT *devtest2006* as validation set and *test2006* as test set. For MultiUN, four languages are selected, Arabic (Ar), Chinese (Zh), Russian (Ru), and English (En). The selected languages are distributed in various language families, making the zero-shot language transfer more difficult. We use MultiUN standard validation and test sets to report the zero-shot performance. The detailed dataset statistics is presented in Appendix. For each dataset, we lowercase all data and preprocess the corpus with 40K BPE operations on all languages. To differentiate language pairs, we follow Johnson et al. (2017) to add a language tag "<2Y>" on the source side for translating $X \Rightarrow Y$.

Experimental Details. We choose standard Transformer-base (Vaswani et al., 2017) architecture to conduct experiments on all baseline and proposed methods, with $n_{\text{layer}} = 6, n_{\text{head}} = 8, d_{\text{embd}} = 512$. We use Fairseq toolkit (Ott et al., 2019) for fast implementations and experiments. We deploy Adam (Kingma and Ba, 2014) ($\beta_1 = 0.9, \beta_2 = 0.98$) optimizer and train the model with $lr = 0.0005, t_{\text{warmup}} = 4000$, dropout $= 0.1, n_{\text{batch}} = 8000$ tokens. Every model is trained for 300k updates (additional 100k for pre-training), and the best model is selected based on BLEU score on validation set every 10k updates. For decoding, we adopt beam-search with beam size = 5.

Baselines. In our experiments, we compare the proposed method **MNMT+DN** with the following approaches: (*i*) **MNMT** (Johnson et al., 2017): training a multilingual NMT model on all directions with available parallel data; (*ii*) **LM+MNMT** (Gu et al., 2019): pre-training the decoder as a multilingual language model, then training the MNMT model initialized with the pre-trained decoder; (*iii*) **MNMT-RC** (Liu et al., 2020): removing residual connections in an encoder layer to disentangle positional information. We re-implement all baseline methods, following the same experimental settings to make fair comparison with our method.

4.2 Results on MultiUN Dataset

Table 2 reports the main results on MultiUN dataset. We can see that our proposed method achieves state-of-the-art performance on all six zero-shot translation directions among all multilingual NMT systems. Also, our method significantly improves the zero-shot performance of vanilla MNMT model by an average 14.4 BLEU score without performance

MultiUN	Ar, Zh, Ru ↔ En							
Model	Ar-Ru		Ar-Zh		Ru-Zh		Zero	Parallel
	←	→	←	→	←	→	Avg.	Avg.
MNMT	17.9	13.4	16.1	29.5	12.1	30.3	19.9	49.2
LM+MNMT	22.0	29.3	20.3	42.7	24.3	42.1	30.1	48.9
MNMT-RS	20.8	26.1	20.3	37.9	24.2	37.4	27.8	49.9
MNMT+DN (Ours)	**24.6**	**33.0**	**24.6**	**47.2**	**30.0**	**46.1**	**34.3**	**50.1**

Table 2: Overall BLEU score [%] on six zero-shot directions of MultiUN dataset. Zero Avg. and Parallel Avg. refer to average BLEU score of six zero-shot directions and six supervised directions, respectively.

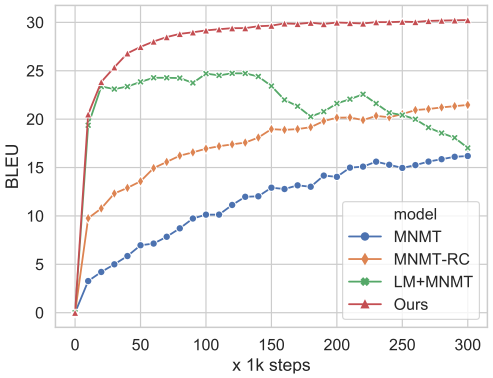

Figure 1: Learning curve of different methods on MultiUN dataset. We sub-sample 1K sentences from every zero-shot translation direction and report BLEU score on the combined 6K-size validation set.

Europarl	De, Fr ↔ En			
Model	De-Fr		Zero	Parallel
	←	→	Avg.	Avg.
MNMT	21.5	27.3	24.4	34.1
LM+MNMT	25.5	31.1	28.3	33.6
MNMT-RC	25.1	30.8	28.0	33.5
MNMT+DN (Ours)	**27.1**	**31.8**	**29.5**	**33.7**

Table 3: Overall BLEU score [%] on two zero-shot directions of Europal dataset.

Europarl	De, Fr ↔ En			
Setting	De-Fr		Zero	Parallel
	←	→	Avg.	Avg.
DE⇒EN+EN⇒FR	-	6.0	-	-
DE⇒EN+EN⇒FR+DN	-	31.1	-	-
MNMT	21.5	27.3	24.4	34.1
BART-PT+MNMT	25.7	31.2	28.5	33.6
MNMT+DN (Ours)	**27.1**	**31.8**	**29.5**	**33.7**

Table 4: BLEU scores [%] of the ablation study on Europarl dataset. "+DN" means that the experiment setting includes denoising autoencoder task.

degradation on supervised directions. These results demonstrate the effectiveness of incorporating denoising autoencoder objective in the training of multilingual NMT. We further investigate the learning curve of different methods on the validation set. As shown in Figure 1, our proposed method reaches faster convergence than MNMT and MNMT-RC, while LM+MNMT easily leads to over-fitting.

4.3 Results on Europarl Dataset

The main results on Europarl dataset are presented in Table 3. We can observe that our proposed method still significantly improves the zero-shot translation performance of multilingual NMT systems with an average 5.1 BLEU score improvements. Different from the MultiUN dataset with four languages distributed in different language families, the selected languages (De, Fr, En) of Europarl are all European languages, making the gap between various baselines and our method smaller than that of MultiUN.

4.4 Ablation Study

To further evaluate the effectiveness of denoising autoencoder task, we conduct an ablation study on Europal dataset as shown in Table 4. With simply incorporating denoising autoencoder task into translation task on only two directions, the model achieves a remarkable zero-shot performance on DE⇒FR of 31.1 BLEU score. The introduction of denoising autoencoder task can effectively break the spurious correlations between output language and semantics, enabling the failed model to perform zero-shot translation. Combined with more translation tasks, MNMT+DN further improve translation accuracy on DE⇒FR. Actually, an alternative to our proposed method is BART pre-training, which learns the denoising autoencoder objective first. We can observe that BART-PT+MNMT gains a similar performance to LM+MNMT, but worse than MNMT+DN due to the catastrophic forgetting problem (McCloskey and Cohen, 1989).

5 Conclusion

In this paper, to improve the zero-shot performance of multilingual NMT system, we proposed to introduce denoising autoencoder objective into conventional translation objective. We analyze the motivation and effectiveness of proposed method from the perspective of latent variables. The experimental results demonstrate that the proposed methods can significantly resolve spurious correlation issue in multilingual NMT and achieves state-of-the-art performance on zero-shot translation.

References

N. Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, M. Johnson, and Wolfgang Macherey. 2019a. The missing ingredient in zero-shot neural machine translation. *ArXiv*, abs/1903.07091.

N. Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, M. Johnson, M. Krikun, M. Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Z. Chen, and Y. Wu. 2019b. Massively multilingual neural machine translation in the wild: Findings and challenges. *ArXiv*, abs/1907.05019.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073*.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. *arXiv preprint arXiv:1906.01181*.

Thanh-Le Ha, J. Niehues, and A. Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *ArXiv*, abs/1611.04798.

Baijun Ji, Zhirui Zhang, Xiangyu Duan, Min Zhang, Boxing Chen, and Weihua Luo. 2020. Cross-lingual pre-training based transfer for zero-shot neural machine translation. *ArXiv*, abs/1912.01214.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, N. Arivazhagan, and Orhan Firat. 2019. Investigating multilingual nmt representations at scale. *ArXiv*, abs/1909.02197.

M. Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *ArXiv*, abs/1910.13461.

Danni Liu, J. Niehues, James Cross, Francisco Guzmán, and X. Li. 2020. Improving zero-shot translation by disentangling positional information. *ArXiv*, abs/2012.15127.

M. McCloskey and N. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165.

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

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

Shuo Ren, Wenhu Chen, Shujie Liu, Mu Li, M. Zhou, and S. Ma. 2018. Triangular architecture for rare language translation. *ArXiv*, abs/1805.04813.

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 *NIPS*.

Biao Zhang, P. Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. *ArXiv*, abs/2004.11867.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534.