Improving Zero-shot Multilingual Neural Machine Translation by Leveraging Cross-lingual Consistency Regularization

Pengzhi Gao, Liwen Zhang, Zhongjun He, Hua Wu, and Haifeng Wang
Baidu Inc. No. 10, Shangdi 10th Street, Beijing, 100085, China
{gaopengzhi,zhangliwen04,hezhongjun,wu_hua,wanghaifeng}@baidu.com

Abstract

The multilingual neural machine translation (NMT) model has a promising capability of zero-shot translation, where it could directly translate between language pairs unseen during training. For good transfer performance from supervised directions to zero-shot directions, the multilingual NMT model is expected to learn universal representations across different languages. This paper introduces a cross-lingual consistency regularization, CrossConST, to bridge the representation gap among different languages and boost zero-shot translation performance. The theoretical analysis shows that CrossConST implicitly maximizes the probability distribution for zero-shot translation, and the experimental results on both low-resource and high-resource benchmarks show that CrossConST consistently improves the translation performance. The experimental analysis also proves that CrossConST could close the sentence representation gap and better align the representation space. Given the universality and simplicity of CrossConST, we believe it can serve as a strong baseline for future multilingual NMT research.

1 Introduction

The objective of multilingual neural machine translation (NMT) is to construct a single, comprehensive model capable of translating between any pair of languages (Firat et al., 2016; Ha et al., 2016; Gu et al., 2018; Zhang et al., 2020; Fan et al., 2021). This not only benefits low-resource translation (Aharoni et al., 2019), but also enables zero-shot translation (Gu et al., 2019). The success of zero-shot translation depends on the capability of the model to learn language-agnostic representations. The conventional multilingual NMT model (Johnson et al., 2017), however, often struggles with learning the universal representations among different languages (Figure 1 (a)), which leads to poor zero-shot translation performance, particularly compared to the pivot-based methods (Cheng et al., 2017).

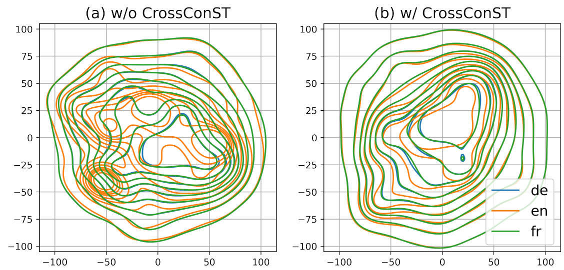

Figure 1: Bivariate kernel density estimation plots of sentence representations after using T-SNE dimensionality reduction on the multi-way parallel testset `newstest2012`, where the max-pooled outputs of the multilingual NMT encoder are applied as the sentence representations. The blue line denotes Germany, the orange line denotes English, and the green line denotes French. This figure shows that the sentence representations are aligned better after utilizing CrossConST.

Several methods have been proposed to improve the zero-shot translation performance by learning language-agnostic representations and maximizing cross-lingual transfer. Some approaches modify the model architecture to achieve universal representations (Lu et al., 2018; Ji et al., 2020; Liu et al., 2021; Chen et al., 2021), while others utilize auxiliary training objectives to encourage similarity between the representations of different languages (Arivazhagan et al., 2019; Al-Shedivat and Parikh, 2019; Pham et al., 2019; Pan et al., 2021). Specifically, Gu and Feng (2022) introduce an agreement-based training approach to help the multilingual NMT model make consistent predictions based on the semantics-equivalent sentences. However, most existing methods are far from being widely used due to the degraded supervised translation performance, complicated algorithm implementation, and tedious hyperparameter search.

In this paper, our primary goal is to provide a simple, easy-to-reproduce, yet effective strategy

Findings of the Association for Computational Linguistics: ACL 2023, pages 12103–12119

for learning multilingual NMT. Inspired by Gao et al. (2022), which boost the NMT performance by leveraging intra-lingual consistency regularization, we here propose a cross-lingual consistency regularization method, CrossConST, to learn the universal representations across different languages (Figure 1 (b)) for boosting the zero-shot translation performance, where we introduce the explicit constraints to the semantic-equivalent sentence pairs by leveraging Kullback-Leibler (KL) regularization. The contributions of this paper can be summarized as follows:

- We propose CrossConST, a simple but effective method with only one hyperparameter for improving the generalization of the multilingual NMT model, and theoretically prove that it implicitly maximizes the probability distribution for zero-shot translation.
- Our experimental results show that CrossConST achieves significant zero-shot translation improvements over the Transformer model on both low-resource and high-resource multilingual translation benchmarks and outperforms the state-of-the-art (SOTA) methods OT & AT (Gu and Feng, 2022) and mRASP2 (Pan et al., 2021) on average.

2 Cross-lingual Consistency for Multilingual NMT

In this section, we formally propose CrossConST, a cross-lingual consistency regularization for learning multilingual NMT. We first review the multilingual neural machine translation (Section 2.1), then introduce our method in detail (Section 2.2). We theoretically analyze the regularization effect of CrossConST (Section 2.3) and propose a two-stage training strategy (Section 2.4).

2.1 Multilingual Neural Machine Translation

Define $L = \{L_1, ..., L_M\}$, where L is a collection of M languages. The multilingual NMT model refers to a neural network with an encoder-decoder architecture, which receives a sentence in language L_i as input and returns a corresponding translated sentence in language L_j as output. Assume $\mathbf{x} = x_1, ..., x_I$ and $\mathbf{y} = y_1, ..., y_J$ that correspond to the source and target sentences with lengths I and J, respectively. Note that x_1 denotes the language identification token to indicate the target language the multilingual NMT model should translate to, and y_J denotes the special end-of-sentence symbol $\langle eos \rangle$. The encoder first maps a source sentence $\mathbf{x}$ into a sequence of word embeddings $e(\mathbf{x}) = e(x_1), ..., e(x_I)$, where $e(\mathbf{x}) \in \mathbb{R}^{d \times I}$, and d is the embedding dimension. The word embeddings are then encoded to the corresponding hidden representations $\mathbf{h}$. Similarly, the decoder maps a shifted copy of the target sentence $\mathbf{y}$, i.e., $\langle bos \rangle, y_1, ..., y_{J-1}$, into a sequence of word embeddings $e(\mathbf{y}) = e(\langle bos \rangle), e(y_1), ..., e(y_{J-1})$, where $\langle bos \rangle$ denotes a special beginning-of-sentence symbol, and $e(\mathbf{y}) \in \mathbb{R}^{d \times J}$. The decoder then acts as a conditional language model that operates on the word embeddings $e(\mathbf{y})$ and the hidden representations $\mathbf{h}$ generated by the encoder.

Let $\mathcal{S}_{i,j}$ denote the parallel corpus of language pair (L_i, L_j), and $\mathcal{S}$ denotes the entire training corpus. The standard training objective is to minimize the empirical risk:

$$\mathcal{L}_{ce}(\theta) = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y}) \in \mathcal{S}} [\ell(f(\mathbf{x}, \mathbf{y}; \theta), \ddot{\mathbf{y}})], \tag{1}$$

where ℓ denotes the cross-entropy loss, θ is a set of model parameters, $f(\mathbf{x}, \mathbf{y}; \theta)$ is a sequence of probability predictions, i.e.,

$$f_j(\mathbf{x}, \mathbf{y}; \theta) = P(y|\mathbf{x}, \mathbf{y}_{<j}; \theta), \tag{2}$$

and $\ddot{\mathbf{y}}$ is a sequence of one-hot label vectors for $\mathbf{y}$.

2.2 CrossConST: A Cross-lingual Consistency Regularization for Multilingual NMT

Consider the multilingual NMT model as a function $f(\mathbf{x}, \mathbf{y}; \theta)$, which could be further decomposed as follows:

$$f(\mathbf{x}, \mathbf{y}; \theta) := f_{dec}(f_{enc}(\mathbf{x}; \theta_{enc}), \mathbf{y}; \theta_{dec}), \tag{3}$$

where $f_{enc}(\cdot)$ and $f_{dec}(\cdot)$ denote the encoder and decoder, and θ_{enc} and θ_{dec} are the sets of parameters for the encoder and decoder respectively. An ideal multilingual NMT model should have the following properties:

- The encoder should output universal representations which are language agnostic. Semantic-equivalent sentences in different languages should share similar representations in the encoder output.
- Given the target language to which the multilingual NMT model should translate to, the decoder should make consistent predictions based on the semantic-equivalent representations in the encoder output.

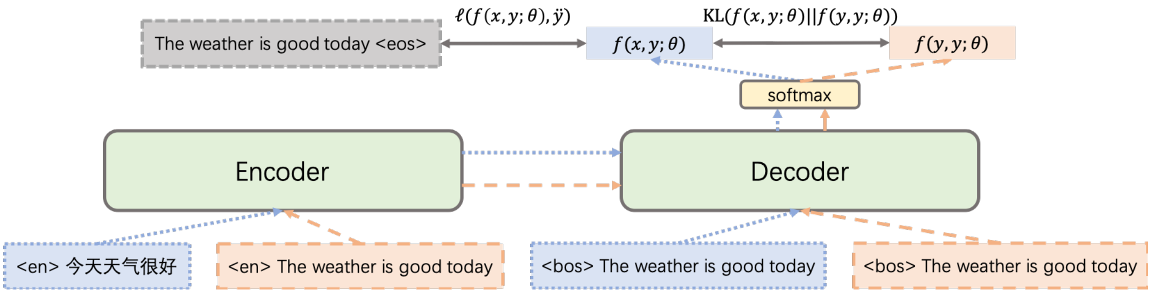

Figure 2: Illustration of the CrossConST regularization, where the original Chinese-English sentence pair ("今天天气很好", "The weather is good today") and the copied English-English sentence pair ("The weather is good today", "The weather is good today") are both go through the multilingual NMT model and obtain two output distributions $f(\mathbf{x}, \mathbf{y}; \theta)$ and $f(\mathbf{y}, \mathbf{y}; \theta)$. The same procedure is also applied to the English-Chinese sentence pair ("The weather is good today", "今天天气很好") during the training of the multilingual NMT model.

The main idea of our method is to close the representation gap among semantic-equivalent sentences in the encoder output and force the output distribution of the decoder to be consistent among different semantic-equivalent representations. During the training of multilingual NMT model, for each sentence pair $(\mathbf{x}, \mathbf{y})$, the training objective of CrossConST is defined as:

$$\mathcal{L}_{CrossConST}(\theta) = \mathcal{L}_{ce}(\theta) + \alpha \mathcal{L}_{kl}(\theta), \quad (4)$$

where

$$\mathcal{L}_{kl}(\theta) = \mathrm{KL}(f(\mathbf{x}, \mathbf{y}; \theta) \| f(\mathbf{y}, \mathbf{y}; \theta)), \quad (5)$$

$\mathrm{KL}(\cdot\|\cdot)$ denotes the Kullback-Leibler (KL) divergence of two distributions, and α is a scalar hyper-parameter that balances $\mathcal{L}_{ce}(\theta)$ and $\mathcal{L}_{kl}(\theta)$. Note that the gradient could be backpropagated through both sides of the KL regularization in CrossConST. Figure 2 illustrates CrossConST regularization for learning multilingual NMT model.

Note that the constraint introduced by (5) forces the equivalence between $f(\mathbf{x}, \mathbf{y}; \theta)$ and $f(\mathbf{y}, \mathbf{y}; \theta)$, which implicitly leads to

$$f_{enc}(\mathbf{x}; \theta_{enc}) = f_{enc}(\mathbf{y}; \theta_{enc}). \quad (6)$$

Semantic-equivalent sentences $\mathbf{x}$ and $\mathbf{y}$ then share similar representations in the encoder output, and the decoder makes consistent predictions based on the semantic-equivalent representations $f_{enc}(\mathbf{x}; \theta_{enc})$ and $f_{enc}(\mathbf{y}; \theta_{enc})$. The properties of the ideal multilingual NMT model implicitly hold.

2.3 Theoretical Analysis

Consider training a multilingual NMT model on the English-centric dataset, where $\mathbf{x}$ and $\mathbf{y}$ denote the sentences in two non-English languages, and $\mathbf{z}$ denotes the English sentence. Let's consider the zero-shot translation direction $\mathbf{x} \rightarrow \mathbf{y}$. Inspired by Ren et al. (2018) and Wang et al. (2021), we take a different approach to modeling the translation probability $P(\mathbf{y}|\mathbf{x}; \theta)$. We introduce language $\mathbf{z}$ as a bridge to connect $\mathbf{x}$ and $\mathbf{y}$. Following Jensen's Inequality, we could derive the lower bound of $P(\mathbf{y}|\mathbf{x}; \theta)$ over the parallel corpus $\mathcal{S}$ as follows:

$$\begin{aligned} \mathcal{L}(\theta) &= \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}} \log P(\mathbf{y}|\mathbf{x}; \theta) \\ &\geq \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}} \sum_{\mathbf{z}} Q(\mathbf{z}; \theta) \log \frac{P(\mathbf{y}|\mathbf{z}; \theta) P(\mathbf{z}|\mathbf{x}; \theta)}{Q(\mathbf{z}; \theta)} \\ &:= \bar{\mathcal{L}}(\theta), \end{aligned}$$

and the gap between $\mathcal{L}(\theta)$ and $\bar{\mathcal{L}}(\theta)$ could be calculated as follows:

$$\begin{aligned} \mathcal{L}(\theta) - \bar{\mathcal{L}}(\theta) &= \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}} \sum_{\mathbf{z}} Q(\mathbf{z}; \theta) \log \frac{Q(\mathbf{z}; \theta)}{P(\mathbf{z}|\mathbf{y}; \theta)} \\ &= \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}} \mathrm{KL}(Q(\mathbf{z}; \theta) \| P(\mathbf{z}|\mathbf{y}; \theta)), \end{aligned}$$

where $Q(\mathbf{z}; \theta)$ is an arbitrary posterior distribution of $\mathbf{z}$. Note that we utilize the approximation that $P(\mathbf{y}|\mathbf{x}, \mathbf{z}; \theta) \approx P(\mathbf{y}|\mathbf{z}; \theta)$ and $P(\mathbf{z}|\mathbf{x}, \mathbf{y}; \theta) \approx P(\mathbf{z}|\mathbf{y}; \theta)$ due to the semantic equivalence of parallel sentences $\mathbf{x}$ and $\mathbf{y}$.

We then introduce the autoencoding task of $\mathbf{z}$ by replacing $Q(\mathbf{z}; \theta)$ with $P(\mathbf{z}|\mathbf{z}; \theta)$ such that

$$\begin{aligned} \bar{\mathcal{L}}(\theta) = \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}} & \mathop{\mathbb{E}}_{\mathbf{z}\sim P(\mathbf{z}|\mathbf{z};\theta)} \log P(\mathbf{y}|\mathbf{z}; \theta) \\ & - \mathrm{KL}(P(\mathbf{z}|\mathbf{z}; \theta) \| P(\mathbf{z}|\mathbf{x}; \theta)) \end{aligned} \quad (7)$$

and

$$\mathcal{L}(\theta) - \bar{\mathcal{L}}(\theta) = \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}} \mathrm{KL}(P(\mathbf{z}|\mathbf{z};\theta)\|P(\mathbf{z}|\mathbf{y};\theta)). \tag{8}$$

To maximize $\mathcal{L}(\theta)$, we should maximize the lower bound $\bar{\mathcal{L}}(\theta)$ and minimize the the gap between $\mathcal{L}(\theta)$ and $\bar{\mathcal{L}}(\theta)$. By utilizing the cross-lingual consistency regularization, CrossConST helps minimize the KL terms in (7) and (8) and implicitly maximizes the probability distributions for zero-shot translation, which results in better translation performance in $\mathbf{x} \rightarrow \mathbf{y}$ direction. The detailed proof can be found in Appendix A.

2.4 Training Strategy: Multilingual NMT Pretraining and CrossConST Finetuning

Inspired by Johnson et al. (2017) and Wu et al. (2021), we only use one language tag to indicate the target language the multilingual NMT model should translate to. For instance, the following English to German sentence pair "How are you? → Wie geht es dir?" is transformed to "<de> How are you? → Wie geht es dir?". And Wu et al. (2021) demonstrate that such language tag strategy could enhance the consistency of semantic representations and alleviate the off-target issue in zero-shot translation directions.

To stabilize the multilingual NMT training procedure and accelerate the convergence of the multilingual NMT model, we adopt a two-stage training strategy. We first train a conventional multilingual NMT model as the pretrained model and then finetune the model with CrossConST objective function (4). It is worth mentioning that Pham et al. (2019) derive a similar problem formulation and training strategy. However, they do not demonstrate the effectiveness of their proposed method (KL Softmax) in Pham et al. (2019). To the best of our knowledge, we for the first time show the effectiveness of the simple cross-lingual consistency regularization for improving the translation performance of the multilingual NMT model. Note that while Pham et al. (2019) decouple the gradient path in the decoder from the KL divergence term, our design allows for backpropagation through both sides of the KL regularization in CrossConST. We do not decouple any gradient path in our model.

3 Low Resource Scenario

We here investigate the performance of CrossConST on the low-resource multilingual machine translation benchmark. For fair comparisons, we keep our experimental settings consistent with the previous work (Gu and Feng, 2022).

3.1 Dataset Description

We conduct our experiments on the IWSLT17 benchmark (Cettolo et al., 2017), which releases a multilingual corpus in five languages: English (`en`), German (`de`), Dutch (`nl`), Romanian (`ro`), and Italian (`it`). We consider the English-centric scenario, where we collect the parallel sentences from/to English. The detailed information of the training dataset is summarized in Table 5 in Appendix B. There are eight supervised translation directions and twelve zero-shot translation directions, and we use the official validation and test sets in our experiments. Following the common practice, we tokenize each language by applying the Moses toolkit (Koehn et al., 2007) and build a shared dictionary with 32K byte-pair-encoding (BPE) (Sennrich et al., 2016) types.

3.2 Model Configuration

We implement our approach on top of the Transformer (Vaswani et al., 2017). We apply a standard base Transformer with 6 encoder and decoder layers, 8 attention heads, embedding size 512, and FFN layer dimension 2048. We apply cross-entropy loss with label smoothing rate 0.1 and set max tokens per batch to be 4096. We use the Adam optimizer with Beta $(0.9, 0.98)$, 4000 warmup updates, and inverse square root learning rate scheduler with initial learning rates $7e^{-4}$. We use dropout rate 0.3 and beam search decoding with beam size 5 and length penalty 0.6. We apply the same training configurations in both pretraining and finetuning stages. We fix α to be 0.25 in (4) for CrossConST. We use case-sensitive sacreBLEU (Post, 2018) to evaluate the translation quality. We train all models until convergence on eight NVIDIA Tesla V100 GPUs. All reported BLEU scores are from a single model. For all the experiments below, we select the saved model state with the best validation performance.

3.3 Main Results

We compare our approach with the following methods on the IWSLT17 benchmark:

- **m-Transformer** (Johnson et al., 2017): A multilingual NMT model that directly learns the

Method	de ↔ it	de ↔ nl	de ↔ ro	it ↔ ro	it ↔ nl	nl ↔ ro	Zero-shot Average	Supervised Average
Pivot†	18.10	19.66	16.49	21.37	21.44	18.70	19.29	-
m-Transformer†	15.46	18.30	14.70	19.03	18.48	16.11	17.01	30.62
SR Alignment†	16.45	18.80	15.45	20.02	19.20	17.25	17.85	30.41
KL-Softmax†	16.06	18.27	15.00	20.09	18.89	16.52	17.46	30.50
mRASP2 w/o AA†	16.98	19.60	15.88	20.75	19.40	17.59	18.36	30.39
DisPos†	16.13	19.21	15.52	20.12	19.58	17.32	17.97	30.49
DAE Training†	16.32	18.69	15.72	20.42	19.11	17.22	17.91	30.51
TGP†	17.64	15.85	16.86	19.34	19.53	20.05	18.21	30.66
LM Pretraining†	17.66	15.86	16.16	19.05	19.02	**20.07**	17.96	30.52
OT & AT†	17.28	19.81	16.09	20.83	20.14	17.85	18.66	30.52
Pivot	18.87	20.09	17.20	21.56	22.22	19.35	19.88	-
OT & AT	18.18	20.22	16.82	21.96	21.15	18.66	19.50	31.14
m-Transformer	17.2	19.61	15.88	20.81	20.21	17.89	18.60	31.34
+ CrossConST	**18.70**	**20.32**	**16.98**	**22.17**	**21.83**	19.30	**19.88**	**31.37**

Table 1: Performance on the IWSLT17 multilingual translation benchmark. Each entry in the first six columns denotes the averaged BLEU scores of both directions. † denotes the numbers are reported from Gu and Feng (2022), others are based on our runs. The highest scores are marked in bold for all models except for the pivot translation in each column. The detailed evaluation results are summarized in Table 8 in Appendix C.

many-to-many translation on the English-centric dataset.

- **Pivot Translation** (Cheng et al., 2017): m-Transformer first translates the source language into English before generating the target language.
- **Sentence Representation Alignment (SR Alignment)** (Arivazhagan et al., 2019): An additional regularization loss is utilized to minimize the discrepancy of the source and target sentence representations.
- **Softmax Forcing (KL-Softmax)** (Pham et al., 2019): This method forces the decoder to generate the target sentence from itself by introducing a KL divergence loss.
- **Contrastive Learning (mRASP2 w/o AA)** (Pan et al., 2021): This method introduces a contrastive loss to minimize the representation gap between the similar sentences and maximize that between the irrelevant sentences. Note that the aligned augmentation (AA) method is not utilized.
- **Disentangling Positional Information (DisPos)** (Liu et al., 2021): This method drops the residual connections in a middle layer of the encoder to achieve the language-agnostic representations.
- **Denosing Training (DAE Training)** (Wang et al., 2021): This approach introduces a denoising autoencoding task during the multilingual NMT model training.
- **Target Gradient Projection (TGP)** (Yang et al., 2021b): This method guides the training with constructed oracle data, where the gradient is projected not to conflict with the oracle gradient.
- **Language Model Pretraining (LM Pretraining)** (Gu et al., 2019): This approach strengthens the decoder language model (LM) prior to NMT model training.
- **Optimal Transport & Agreement-based Training (OT & AT)** (Gu and Feng, 2022): This method proposes an optimal transport loss to bridge the gap between the semantic-equivalent representations and an agreement-based loss to force the decoder to make consistent predictions based on semantic-equivalent sentences. We set γ_1 and γ_2 in OT & AT to be 0.4 and 0.001 respectively in the experiments.

We report test BLEU scores of all comparison methods and our approach on the IWSLT17 dataset in Table 1. We can see that our multilingual NMT model achieves strong or SOTA BLEU scores in both supervised and zero-shot translation directions. Note that our approach outperforms OT & AT even though its implementation is much more complicated than ours. It is worth mentioning that CrossConST is the only method that can achieve a similar zero-shot translation performance compared with the pivot translation. Note that the BLEU scores of our m-Transformer, especially in the zero-shot translation directions, are higher than that reported in Gu and Feng (2022). Such gap might be due to the different language tag strategies used in Gu and Feng (2022) and our experiments, which is in line with Wu et al. (2021).

3.4 Does CrossConST Still Work Beyond English-centric Scenario?

We here extend our experiments on the IWSLT17 benchmark beyond the English-centric scenario. Specifically, we gather the English-centric dataset used in Section 3.3 and supplement it with an additional 20K `de` ↔ `it` sentence pairs, which are subsampled from the IWSLT17 dataset. This experimental setup is highly practical because the size of the non-English datasets is usually an order less than that of the English-centric dataset.

Method	Training Dataset	Zero-shot Average	Supervised Average
m-Transformer	①	18.60	31.34
+ CrossConST	①	19.88	31.37
m-Transformer	②	19.76	31.59
+ CrossConST	②	**20.35**	**31.67**

Table 2: Performance on the IWSLT17 multilingual translation benchmark. ① denotes the English-centric dataset. ② denotes the English-centric dataset + extra `de` ↔ `it` dataset. The detailed evaluation results are summarized in Table 9 in Appendix C.

We report test BLEU scores of the baseline and our approach on the IWSLT17 dataset in Table 2. By checking model performance under different combinations of dataset and training strategy, we have the following observations: 1) Adding beyond the English-centric dataset (`de` ↔ `it`) could greatly improve the overall zero-shot translation performance. 2) The CrossConST is complementary to the data-based method and could further improve the performance of the zero-shot translation.

4 High Resource Scenario

We here investigate the performance of the CrossConST on the high-resource multilingual machine translation benchmark. For fair comparisons, we keep our experimental settings consistent with the previous works (Lin et al., 2020; Pan et al., 2021).

4.1 Dataset Description

We conduct our experiments on PC32, a multilingual parallel corpus of 32 English-centric language pairs. We collect the pre-processed PC32 dataset from Lin et al. (2020)'s release[1]. We also collect the pre-processed PC32 dataset after applying random aligned substitution (RAS) technique from Lin et al. (2020)'s release. The detailed statistics of all training datasets are summarized in Tables 6 and 7 in Appendix B.

For supervised directions, we collect testsets from WMT benchmarks, where four languages, Spanish (`es`), Finnish (`fi`), French (`fr`), and Turkish (`tr`), are selected, resulting in 8 translation directions. We use `multi-bleu.pl`[2] for tokenized BLEU (Papineni et al., 2002) evaluation, where both reference and hypothesis are tokenized by Sacremoses[3]. For zero-shot directions, we collect OPUS-100 zero-shot testsets from Zhang et al. (2020)'s release[4], where six languages, Arabic (`ar`), German (`de`), French (`fr`), Dutch (`nl`), Russian (`ru`), and Chinese (`zh`), are selected, resulting in 25 translation directions. Note that Dutch is not covered in our training dataset such that we only evaluate the zero-shot directions when Dutch is at the source side. We evaluate the multilingual NMT models by case-sensitive sacreBLEU (Post, 2018).

4.2 Model Configuration

We apply a Transformer with 12 encoder and decoder layers, 16 attention heads, embedding size 1024, and FFN layer dimension 4096. We use dropout rate 0.1, learning rate $3e^{-4}$ with polynomial decay scheduling and 10000 warmup updates. We use Adam optimizer with Beta $(0.9, 0.98)$ and $\epsilon = 1e^{-6}$. We set the threshold of gradient norm to be 5.0. We apply cross-entropy loss with label smoothing rate 0.1 and set max tokens per batch to be 1536 with update frequency 50. We use beam search decoding with beam size 5 and length penalty 1.0. We apply the same training configurations in both pretraining and finetuning stages. We fix α to be 0.1 in (4) for CrossConST. We train all models until convergence on 8×4 NVIDIA Tesla V100 GPUs. All reported BLEU scores are from a single model. We select the saved model state with the best validation performance for all the experiments below.

4.3 Main Results

We compare our approach with the following methods on the PC32 benchmark:

[1]https://github.com/linzehui/mRASP

[2]https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

[3]https://github.com/alvations/sacremoses

[4]https://opus.nlpl.eu/opus-100.php

Method	Training Dataset	en - fr WMT14 →	←	en - tr WMT17 →	←	en - es WMT13 →	←	en - fi WMT17 →	←	Average
m-Transformer†	①	42.0	38.1	18.8	23.1	32.8	33.7	20.0	28.2	29.66
mRASP2 w/o AA†	①	42.1	38.7	18.2	24.8	33.1	33.2	20.0	27.8	29.74
mRASP†	②	43.1	39.2	20.0	25.2	34.0	34.3	22.0	29.2	30.88
mRASP2 w/o MC24†	②	43.3	39.3	20.4	**25.7**	34.1	**34.3**	22.0	29.4	31.06
mRASP2†	③	43.5	39.3	21.4	25.8	34.5	35.0	23.4	30.1	31.63
m-Transformer	①	43.5	40.3	20.8	23.8	33.4	32.7	22.0	28.8	30.66
+ CrossConST	①	44.1	**40.7**	21.2	24.5	33.8	33.0	22.2	29.5	31.13
mRASP	②	44.5	39.7	22.1	23.6	33.9	33.1	23.3	29.0	31.15
+ CrossConST	②	**44.6**	**40.7**	**22.4**	24.4	**34.3**	33.7	**23.5**	**29.7**	**31.66**

Table 3: Performance (tokenized BLEU) on WMT supervised translation directions. † denotes the numbers are reported from Pan et al. (2021), others are based on our runs. The highest scores are marked in bold for all models except for mRASP2 in each column. ① denotes PC32. ② denotes PC32 + RAS. ③ denotes PC32 + RAS + MC24.

Method	x - ar →	←	x - zh →	←	x - nl* ←	x - fr →	←	x - de →	←	x - ru →	←	Avg.
Pivot†	5.5	21.1	28.5	20.3	6.0	26.1	23.9	14.4	16.6	16.6	24.6	18.22
m-Transformer†	3.7	6.7	6.7	5.0	6.3	7.7	5.0	4.2	4.9	5.7	5.6	5.60
mRASP2 w/o AA†	4.8	17.1	26.1	15.8	6.4	22.9	21.2	11.8	15.3	13.3	21.4	15.79
mRASP†	4.1	4.4	8.2	4.0	5.1	2.4	7.6	6.2	4.1	4.1	4.6	4.97
mRASP2 w/o MC24†	**5.9**	**18.3**	**27.5**	16.5	**9.6**	**25.2**	21.6	11.2	**16.7**	15.6	**21.7**	17.07
mRASP2†	5.3	20.8	29.0	17.7	6.1	23.6	23.0	12.3	16.4	16.4	22.8	17.32
Pivot (m-Transformer)	6.6	22.2	29.5	21.4	8.7	27.5	24.7	15.7	17.1	18.0	25.3	19.46
Pivot (mRASP)	6.9	21.9	29.4	21.8	8.1	27.2	25.3	15.5	17.2	18.3	25.6	19.49
m-Transformer	5.3	11.2	17.4	16.5	7.5	16.8	21.3	9.8	13.1	14.5	8.2	12.75
+ CrossConST	5.4	17.7	27.2	18.4	9.3	24.0	23.9	14.0	16.0	15.9	20.5	17.30
mRASP	5.6	13.7	24.1	18.3	7.2	17.7	23.0	11.1	13.1	15.5	15.5	14.80
+ CrossConST	**5.9**	16.7	27.2	**19.6**	9.2	23.5	**24.6**	**14.3**	16.0	**16.4**	20.9	**17.48**

Table 4: Performance (de-tokenized BLEU using SacreBLEU) on OPUS-100 zero-shot translation directions. † denotes the numbers are reported from Pan et al. (2021), others are based on our runs. ∗ indicates that Dutch (`nl`) is not included in PC32. The highest scores are marked in bold for all models except for the pivot translation and mRASP2 in each column.

- **mRASP** (Lin et al., 2020): This method proposes a random aligned substitution (RAS) technique that builds code-switched sentence pairs for multilingual pretraining. Note that the results of mRASP reported in this paper are obtained without finetuning.

- **mRASP2** (Pan et al., 2021): This method utilizes the RAS technique on both the bilingual dataset (PC32) and an additional monolingual dataset (MC24). It introduces a contrastive loss to minimize the representation gap between the similar sentences and maximize that between the irrelevant sentences. mRASP2 w/o AA only adopts the contrastive loss based on m-Transformer, and mRASP2 w/o MC24 excludes MC24 from mRASP2.

We report test BLEU scores of all comparison methods and our approach on WMT supervised translation directions in Table 3. With CrossConST regularization, our multilingual NMT model achieves strong or SOTA BLEU scores on the supervised translation directions. Note that all comparison methods and our approach share the same model architecture, and the only differences are the training dataset and the objective loss function. We report test BLEU scores of all comparison methods and our approach on OPUS-100 zero-shot translation directions in Table 4, which includes six languages and 25 translation directions in total. The detailed evaluation results are summarized in Table 10 in Appendix D. We also report the evaluation results of the pivot translation based on m-Transformer and mRASP. We can see that CrossConST greatly boosts the performance in the zero-shot translation directions and substantially narrows the performance gap with the pivot translation. It is worth mentioning that our approach could improve zero-shot translation by a large margin and also benefit the supervised translation.

By checking model performance under different scenarios, we have the following observations: 1) Our language tag strategy works better than that in Pan et al. (2021) for learning the multilingual NMT model on the English-centric dataset, especially for the zero-shot translation, which is in line with Wu et al. (2021). 2) CrossConST is crucial for the performance improvement in the zero-shot translation directions and performs slightly better when combined with the code-switched training dataset. 3) Our approach could outperform mRASP2 on average in the absence of the MC24 dataset, which implies the effectiveness of CrossConST compared with the contrastive loss utilized in mRASP2.

4.4 Does CrossConST Really Learn A Better Latent Representation?

We conduct the experiments on the multi-way parallel testset `newstest2012`[5] from the WMT13 (Bojar et al., 2013) translation task, where 3003 sentences have translations in six languages: Czech (`cs`), Germany (`de`), English (`en`), Spanish (`es`), French (`fr`), and Russian (`ru`). We calculate the sentence representations by max-pooling the multilingual NMT encoder outputs.

Sentence Representation Visualization To verify whether CrossConst can better align different languages' semantic space, we visualize the sentence representations of Germany (`de`), English (`en`), and French (`fr`). We apply dimension reduction on the 1024-dimensional sentence representations with T-SNE (Hinton and Roweis, 2002) and then depict the bivariate kernel density estimation based on the 2-dimensional representations in Figure 1. Figure 1 shows that m-Transformer cannot align these three languages well in the representation space, while CrossConST draws the sentence representations across different languages much closer. Please check Figures 4 and 5 in Appendix E for the visualization of the sentence representations in other languages.

Multilingual Similarity Search We conduct the multilingual similarity search experiment to verify that CrossConST indeed closes the latent representation gap among different languages. For each sentence in the source language, we find the closest sentence in the target language according to the cosine similarity of the corresponding sentence representations. The evaluation results are reported in Figure 3. By checking model performance on different language pairs, we have the following observations: 1) m-Transformer could achieve decent performance (94.71% on average) among non-English directions. However, the similarity search accuracy degrades dramatically (81.03% on average) in the English-centric directions, which implies that English does not align well with non-English languages in m-Transformer. We think such bad representation alignment between English and non-English languages is one of the critical reasons that m-Transformer underperforms in the zero-shot translation directions compared with the pivot-based method. 2) CrossConST significantly improves the similarity search performance in the English-centric direction (14.74% improvement on average) and further boosts the performance among non-English directions (1% improvement on aver-

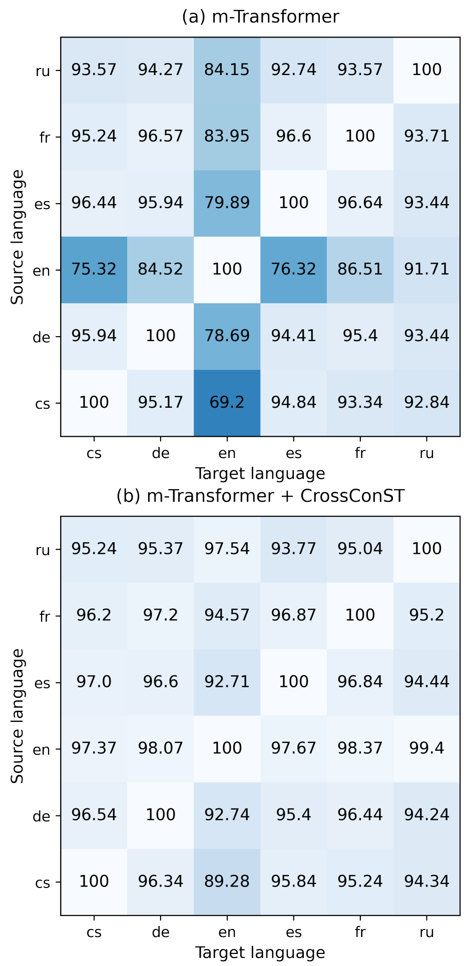

Figure 3: Similarity search accuracy of m-Transformer with/without CrossConST for different language pairs.

[5]https://www.statmt.org/wmt13/dev.tgz

age). We believe the improvement of similarity search accuracy could be regarded as an indicator of better cross-lingual representation alignment and confirm that CrossConST can learn effective universal representation across different languages.

5 Related Work

Early works on multilingual NMT demonstrate its zero-shot translation capability (Ha et al., 2016; Johnson et al., 2017). To further improve the zero-shot translation performance, one direction is to force the multilingual NMT encoder output to be language-agnostic via additional regularization constraints or training tasks (Pham et al., 2019; Arivazhagan et al., 2019; Wei et al., 2020; Liu et al., 2021; Wang et al., 2021; Yang et al., 2021b; Gu and Feng, 2022). For example, Gu and Feng (2022) introduce an agreement-based training approach to help the multilingual NMT model make consistent predictions based on the semantics-equivalent sentences. Our method follows this line but outperforms these methods by introducing a simple yet effective cross-lingual regularization constraint, which effectively reduces discrepancies in representations across languages.

Another direction is to utilize extra data such as generated pseudo sentence pairs, monolingual datasets, and pretrained models (Gu et al., 2019; Al-Shedivat and Parikh, 2019; Zhang et al., 2020; Chen et al., 2021; Yang et al., 2021a). For example, Al-Shedivat and Parikh (2019) encourages the multilingual NMT model to produce equivalent translations of parallel training sentence pairs into an auxiliary language. Zhang et al. (2020) proposes random online back-translation to enforce the translation of unseen training language pairs. Unlike these methods, CrossConST does not require additional data and is orthogonal to these methods. We could further boost the zero-shot translation performance by combining our method with these data-driven approaches.

6 Conclusion

In this paper, we propose CrossConST: a simple but effective cross-lingual consistency regularization method for learning multilingual NMT models. We theoretically analyze the regularization effect of CrossConST and verify its effectiveness for zero-shot translation. For the stable training of multilingual NMT, we propose a two-state training strategy that consists of multilingual NMT pretraining and CrossConST finetuning. Experiments on low and high resource multilingual translation benchmarks demonstrate CrossConST's capabilities to improve translation performance in both supervised and zero-shot directions. Further experimental analysis confirms that our method indeed leads to better cross-lingual representation alignment. Given its universality and simplicity, we anticipate that researchers could leverage the simplicity of CrossConST as a foundation to achieve new SOTA results in their own work. For future work, we will explore the effectiveness of CrossConST on more multilingual tasks, such as multilingual sentence embedding, multilingual word alignment, etc.

Limitations

In this paper, we mainly focus on evaluating our approach on two English-centric corpora, IWSLT17 and PC32. Future research could consider more multilingual machine translation benchmarks with different number of languages and training samples and conduct experiments on more challenging training scenarios such as chain configurations where we have multiple bridge languages and different zero-shot distances.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Maruan Al-Shedivat and Ankur Parikh. 2019. Consistency by agreement in zero-shot 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 1184–1197, Minneapolis, Minnesota. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. The missing ingredient in zero-shot neural machine translation. arXiv preprint arXiv:1903.07091.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In Proceedings of the 14th International Conference on Spoken Language Translation, pages 2–14, Tokyo, Japan. International Workshop on Spoken Language Translation.

Guanhua Chen, Shuming Ma, Yun Chen, Li Dong, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. 2021. Zero-shot cross-lingual transfer of neural machine translation with multilingual pretrained encoders. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 15–26, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yong Cheng, Qian Yang, Yang Liu, Maosong Sun, and Wei Xu. 2017. Joint training for pivot-based neural machine translation. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, pages 3974–3980.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. Journal of Machine Learning Research, 22(107):1–48.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 866–875, San Diego, California. Association for Computational Linguistics.

Pengzhi Gao, Zhongjun He, Hua Wu, and Haifeng Wang. 2022. Bi-SimCut: A simple strategy for boosting neural machine translation. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3938–3948, Seattle, United States. Association for Computational Linguistics.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1258–1268, Florence, Italy. Association for Computational Linguistics.

Shuhao Gu and Yang Feng. 2022. Improving zero-shot multilingual translation with universal representations and cross-mapping. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 6492–6504, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thanh-Le Ha, Jan Niehues, and Alex Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In Proceedings of the 13th International Conference on Spoken Language Translation, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Geoffrey E Hinton and Sam Roweis. 2002. Stochastic neighbor embedding. In Advances in Neural Information Processing Systems, volume 15. MIT Press.

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. Proceedings of the AAAI Conference on Artificial Intelligence, 34(01):115–122.

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, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages

2649–2663, Online. Association for Computational Linguistics.

Danni Liu, Jan Niehues, James Cross, Francisco Guzmán, and Xian Li. 2021. Improving zero-shot translation by disentangling positional information. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1259–1273, Online. Association for Computational Linguistics.

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 84–92, Brussels, Belgium. Association for Computational Linguistics.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 244–258, Online. Association for Computational Linguistics.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, and Alexander Waibel. 2019. Improving zero-shot translation with language-independent constraints. In Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers), pages 13–23, Florence, Italy. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Shuo Ren, Wenhu Chen, Shujie Liu, Mu Li, Ming Zhou, and Shuai Ma. 2018. Triangular architecture for rare language translation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 56–65, Melbourne, Australia. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

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 Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008.

Weizhi Wang, Zhirui Zhang, Yichao Du, Boxing Chen, Jun Xie, and Weihua Luo. 2021. Rethinking zero-shot neural machine translation: From a perspective of latent variables. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4321–4327, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2020. On learning universal representations across languages. arXiv preprint arXiv:2007.15960.

Liwei Wu, Shanbo Cheng, Mingxuan Wang, and Lei Li. 2021. Language tags matter for zero-shot neural machine translation. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3001–3007, Online. Association for Computational Linguistics.

Jian Yang, Yuwei Yin, Shuming Ma, Haoyang Huang, Dongdong Zhang, Zhoujun Li, and Furu Wei. 2021a. Multilingual agreement for multilingual neural machine translation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 233–239, Online. Association for Computational Linguistics.

Yilin Yang, Akiko Eriguchi, Alexandre Muzio, Prasad Tadepalli, Stefan Lee, and Hany Hassan. 2021b. Improving multilingual translation by representation and gradient regularization. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7266–7279, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1628–1639, Online. Association for Computational Linguistics.

A Theoretical Discussion of CrossConST

We first discuss how to derive the lower bound of $\mathcal{L}(\theta)$ as follows. Please note that we drop θ

in the following proofs for the simplicity of the expression.

$$
\begin{aligned}
&\mathcal{L}(\theta) \\
&= \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}} \log P(\mathbf{y}|\mathbf{x}) \\
&= \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}} \log \sum_{\mathbf{z}} P(\mathbf{y}|\mathbf{x},\mathbf{z})P(\mathbf{z}|\mathbf{x}) \\
&\approx \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}} \log \sum_{\mathbf{z}} P(\mathbf{y}|\mathbf{z})P(\mathbf{z}|\mathbf{x}) \\
&= \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}} \log \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{z}) \frac{P(\mathbf{y}|\mathbf{z})P(\mathbf{z}|\mathbf{x})}{P(\mathbf{z}|\mathbf{z})} \\
&\geq \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}} \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{z}) \log \frac{P(\mathbf{y}|\mathbf{z})P(\mathbf{z}|\mathbf{x})}{P(\mathbf{z}|\mathbf{z})} \\
&= \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}} \mathop{\mathbb{E}}_{\mathbf{z}\sim P(\mathbf{z}|\mathbf{z})} \log P(\mathbf{y}|\mathbf{z}) \\
&\quad - \mathrm{KL}(P(\mathbf{z}|\mathbf{z}) \| P(\mathbf{z}|\mathbf{x})) \\
&:= \bar{\mathcal{L}}(\theta)
\end{aligned}
$$

We then discuss how to derive the gap between $\mathcal{L}(\theta)$ and $\bar{\mathcal{L}}(\theta)$ as follows.

$$
\begin{aligned}
&\mathcal{L}(\theta) - \bar{\mathcal{L}}(\theta) \\
&= \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}} \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{z}) \log \frac{P(\mathbf{z}|\mathbf{z})P(\mathbf{y}|\mathbf{x})}{P(\mathbf{y}|\mathbf{z})P(\mathbf{z}|\mathbf{x})} \\
&= \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}} \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{z}) \log \frac{P(\mathbf{z}|\mathbf{z})P(\mathbf{y}|\mathbf{x})P(\mathbf{z}|\mathbf{y})}{P(\mathbf{y}|\mathbf{z})P(\mathbf{z}|\mathbf{x})P(\mathbf{z}|\mathbf{y})} \\
&\approx \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}} \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{z}) \log \frac{P(\mathbf{z}|\mathbf{z})P(\mathbf{y}|\mathbf{x})P(\mathbf{z}|\mathbf{y})}{P(\mathbf{y}|\mathbf{x},\mathbf{z})P(\mathbf{z}|\mathbf{x})P(\mathbf{z}|\mathbf{y})} \\
&= \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}} \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{z}) \log \frac{P(\mathbf{z}|\mathbf{z})P(\mathbf{y}|\mathbf{x})P(\mathbf{z}|\mathbf{y})}{P(\mathbf{y},\mathbf{z}|\mathbf{x})P(\mathbf{z}|\mathbf{y})} \\
&= \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}} \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{z}) \log \frac{P(\mathbf{z}|\mathbf{z})P(\mathbf{y}|\mathbf{x})P(\mathbf{z}|\mathbf{y})}{P(\mathbf{z}|\mathbf{x},\mathbf{y})P(\mathbf{y}|\mathbf{x})P(\mathbf{z}|\mathbf{y})} \\
&\approx \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}} \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{z}) \log \frac{P(\mathbf{z}|\mathbf{z})P(\mathbf{y}|\mathbf{x})P(\mathbf{z}|\mathbf{y})}{P(\mathbf{z}|\mathbf{y})P(\mathbf{y}|\mathbf{x})P(\mathbf{z}|\mathbf{y})} \\
&= \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}} \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{z}) \log \frac{P(\mathbf{z}|\mathbf{z})}{P(\mathbf{z}|\mathbf{y})} \\
&= \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}} \mathrm{KL}(P(\mathbf{z}|\mathbf{z}) \| P(\mathbf{z}|\mathbf{y})),
\end{aligned}
$$

where we utilize two approximations as follows:

$$P(\mathbf{y}|\mathbf{x},\mathbf{z}) \approx P(\mathbf{y}|\mathbf{z}) \tag{9}$$

and

$$P(\mathbf{z}|\mathbf{x},\mathbf{y}) \approx P(\mathbf{z}|\mathbf{y}). \tag{10}$$

B Statistics of all training datasets

en ↔	#sentences	en ↔	#sentences
de	446324	nl	510580
it	501278	ro	477316

Table 5: Statistics of IWSLT17 dataset. Each entry shows the total number of parallel sentence pairs for both directions. Note that en → and en ← directions have the equal number of sentence pairs.

en ↔	#sentences	en ↔	#sentences
af	80616	ja	4146998
ar	2424336	ka	400868
be	51008	kk	246622
bg	6305372	ko	2945682
cs	1639292	lt	4721996
de	9420278	lv	6261224
el	2678292	mn	61200
eo	134972	ms	3273034
es	4228938	mt	354488
et	4579720	my	57076
fi	4113282	ro	1550552
fr	74445068	ru	3686958
gu	22792	sr	269302
he	664818	tr	771426
hi	2699732	vi	6450690
it	4144732	zh	44771930

Table 6: Statistics of PC32 dataset. Each entry shows the total number of parallel sentence pairs for both directions. Note that en → and en ← directions have the equal number of sentence pairs.

C Details of Evaluation Results on IWSLT17

D Details of Evaluation Results on OPUS-100

E Sentence Representation Visualization

en →	#sentences	en ←	#sentences	en →	#sentences	en ←	#sentences
af	58723	af	42429	ja	2989787	ja	2072284
ar	1786139	ar	1212160	ka	281346	ka	200434
be	41052	be	25504	kk	132937	kk	123309
bg	5360004	bg	3152631	ko	2130540	ko	1472841
cs	1455275	cs	819418	lt	3545300	lt	2359916
de	8251292	de	4707481	lv	5179183	lv	3130536
el	2402732	el	1333533	mn	49882	mn	30600
eo	93519	eo	67486	ms	2268324	ms	1636517
es	3787101	es	2111065	mt	306122	mt	177244
et	3289592	et	2289755	my	48497	my	28538
fi	3571662	fi	2054925	ro	1359006	ro	775197
fr	63591612	fr	37222318	ru	2859034	ru	1843417
gu	11868	gu	11395	sr	229641	sr	134651
he	532895	he	332357	tr	660576	tr	385713
hi	1990436	hi	1349767	vi	4542508	vi	3225345
it	3733382	it	2068077	zh	37297105	zh	22385733

Table 7: Statistics of PC32 with RAS dataset. Each entry shows the total number of parallel sentence pairs for each direction.

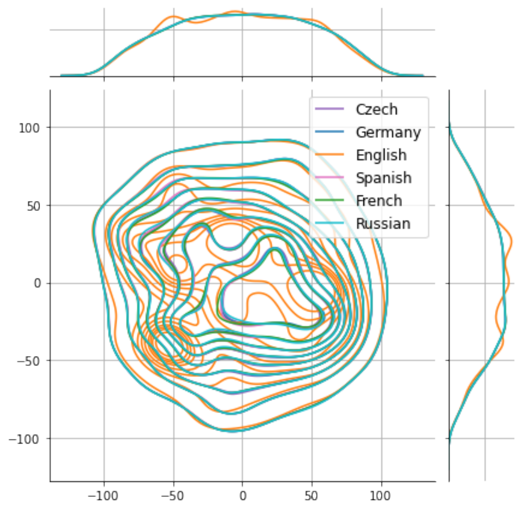

Figure 4: Bivariate kernel density estimation plots of sentence representations based on m-Transformer without CrossConST.

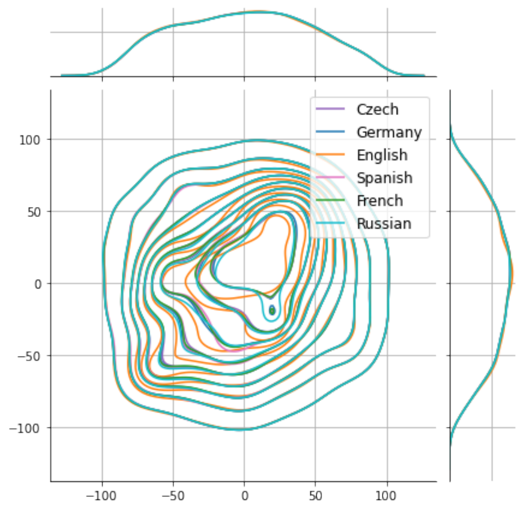

Figure 5: Bivariate kernel density estimation plots of sentence representations based on m-Transformer with CrossConST.

Method	de - it		de - nl		de - ro		it - ro		it - nl	
	→	←	→	←	→	←	→	←	→	←
Pivot	18.81	18.92	19.87	20.3	16.26	18.13	20.19	22.93	22.2	22.23
OT & AT	18.17	18.18	20.17	20.27	16.12	17.52	20.14	23.77	21.07	21.22
m-Transformer	17.18	17.22	19.21	20.01	15.21	16.54	19.27	22.35	20.31	20.1
+ CrossConST	18.6	18.79	20.41	20.22	15.9	18.06	21.02	23.31	21.88	21.77

Method	nl - ro		en - de		en - it		en - nl		en - ro	
	→	←	→	←	→	←	→	←	→	←
Pivot	18.06	20.64	-	-	-	-	-	-	-	-
OT & AT	17.81	19.51	24.87	28.67	35.29	37.61	31.04	33.03	26.17	32.45
m-Transformer	16.65	19.12	24.73	28.49	35.34	38.12	31.64	33.47	26.36	32.56
+ CrossConST	18.21	20.38	24.7	28.87	35.02	38.18	31.75	33.16	26.65	32.66

Table 8: Performance on IWSLT17 supervised and zero-shot translation directions with the English-centric training dataset.

Method	de - it		de - nl		de - ro		it - ro		it - nl	
	→	←	→	←	→	←	→	←	→	←
m-Transformer	18.55	18.88	20.5	20.56	16.06	17.93	20.47	23.42	22.06	21.66
+ CrossConST	19.35	19.63	20.69	20.7	16.48	18.33	21.23	23.74	22.75	22.31

Method	nl - ro		en - de		en - it		en - nl		en - ro	
	→	←	→	←	→	←	→	←	→	←
m-Transformer	17.79	19.28	25.24	29.1	35.42	38.32	31.09	33.32	26.95	33.3
+ CrossConST	18.45	20.57	24.88	29.35	35.46	38.39	31.41	33.38	26.87	33.65

Table 9: Performance on IWSLT17 supervised and zero-shot translation directions with the English-centric and extra de $\leftrightarrow$ it training dataset.

m-Transformer							**m-Transformer + CrossConST**						
	ar	zh	fr	de	ru	Avg		ar	zh	fr	de	ru	Avg
ar→	-	15.8	9.4	6.6	13.0	11.2	ar→	-	27.6	19.1	11.0	13.1	17.7
zh→	6.4	-	33.1	6.6	19.9	16.5	zh→	6.2	-	36.4	9.6	21.3	18.4
fr→	6.8	40.0	-	16.3	22.2	21.3	fr→	7.0	43.6	-	21.1	24.0	23.9
de→	4.2	16.5	18.6	-	13.2	13.1	de→	4.9	19.6	24.4	-	15.2	16.0
ru→	6.6	7.7	9.8	8.5	-	8.2	ru→	5.7	37.7	24.1	14.3	-	20.5
nl→	2.3	6.8	12.9	11.1	4.4	7.5	nl→	3.1	7.6	16.2	13.8	5.9	9.3
Avg	5.3	17.4	16.8	9.8	14.5	**12.75**	Avg	5.4	27.2	24.0	14.0	15.9	**17.30**

mRASP							**mRASP + CrossConST**						
	ar	zh	fr	de	ru	Avg		ar	zh	fr	de	ru	Avg
ar→	-	22.6	10.6	7.7	13.7	13.7	ar→	-	26.2	16.0	11.0	13.4	16.7
zh→	7.1	-	35.2	9.4	21.6	18.3	zh→	6.7	-	37.3	11.6	22.9	19.6
fr→	7.4	41.9	-	18.7	24.1	23.0	fr→	7.8	43.9	-	21.6	24.9	24.6
de→	4.0	16.6	17.2	-	14.4	13.1	de→	4.9	19.6	24.3	-	15.3	16.0
ru→	7.2	33.6	11.8	9.3	-	15.5	ru→	6.9	38.8	24.0	13.9	-	20.9
nl→	2.4	5.9	13.6	10.4	3.7	7.2	nl→	3.3	7.5	15.9	13.6	5.7	9.2
Avg	5.6	24.1	17.7	11.1	15.5	**14.80**	Avg	5.9	27.2	23.5	14.3	16.4	**17.48**

Pivot (m-Transformer)							**Pivot (mRASP)**						
	ar	zh	fr	de	ru	Avg		ar	zh	fr	de	ru	Avg
ar→	-	33.0	24.1	14.0	17.7	22.2	ar→	-	32.2	23.1	14.3	17.8	21.9
zh→	8.8	-	38.1	13.2	25.5	21.4	zh→	9.5	-	38.3	13.0	26.3	21.8
fr→	7.9	44.3	-	21.9	24.7	24.7	fr→	8.4	44.5	-	22.5	25.6	25.3
de→	5.2	21.4	25.7	-	16.2	17.1	de→	5.0	21.7	25.6	-	16.6	17.2
ru→	8.1	41.4	34.8	16.8	-	25.3	ru→	8.9	41.8	35.0	16.7	-	25.6
nl→	2.9	7.6	14.8	12.6	5.7	8.7	nl→	2.9	7.0	14.1	11.2	5.2	8.1
Avg	6.6	29.5	27.5	15.7	18.0	**19.46**	Avg	6.9	29.4	27.2	15.5	18.3	**19.49**

Table 10: Performance (de-tokenized BLEU using SacreBLEU) on OPUS100 zero-shot translation directions.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3 and 4

☑ B1. Did you cite the creators of artifacts you used?
3 and 4

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3 and 4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☑ Did you run computational experiments?

3 and 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
3 and 4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
3 and 4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
3 and 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
3 and 4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.