Contrastive Learning for Many-to-many Multilingual Neural Machine Translation

Xiao Pan, Mingxuan Wang, Liwei Wu, Lei Li
ByteDance AI Lab
{panxiao.94,wangmingxuan.89,wuliwei.000,lileilab}@bytedance.com

Abstract

Existing multilingual machine translation approaches mainly focus on English-centric directions, while the non-English directions still lag behind. In this work, we aim to build a many-to-many translation system with an emphasis on the quality of non-English language directions. Our intuition is based on the hypothesis that a universal cross-language representation leads to better multilingual translation performance. To this end, we propose mRASP2, a training method to obtain a single unified multilingual translation model. mRASP2 is empowered by two techniques: *a)* a contrastive learning scheme to close the gap among representations of different languages, and *b)* data augmentation on both multiple parallel and monolingual data to further align token representations. For English-centric directions, mRASP2 outperforms existing best unified model and achieves competitive or even better performance than the pre-trained and fine-tuned model mBART on tens of WMT's translation directions. For non-English directions, mRASP2 achieves an improvement of average 10+ BLEU compared with the multilingual Transformer baseline. Code, data and trained models are available at https://github.com/PANXiao1994/mRASP2.

1 Introduction

Transformer (Vaswani et al., 2017) has achieved decent performance for machine translation with rich bilingual parallel corpora. Recent work on multilingual machine translation aims to create a single unified model to translate many languages (Johnson et al., 2017; Aharoni et al., 2019; Zhang et al., 2020; Fan et al., 2020; Siddhant et al., 2020). Multilingual translation models are appealing for two reasons. First, they are model efficient, enabling easier deployment (Johnson et al., 2017). Further, parameter sharing across different languages encourages knowledge transfer, which benefits low-resource translation directions and potentially enables zero-shot translation (i.e. direct translation between a language pair not seen during training) (Ha et al., 2017; Gu et al., 2019; Ji et al., 2020).

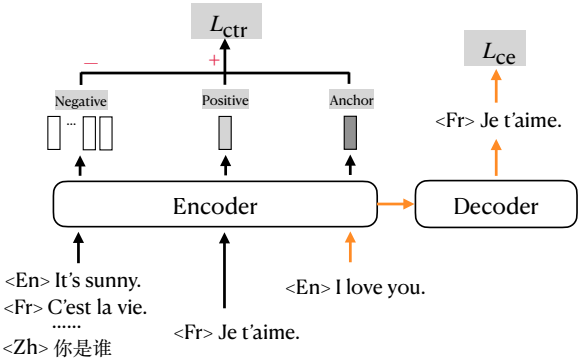


Figure 1: The proposed mRASP2. It takes a pair of parallel sentences (or augmented pseudo-pair) and computes normal cross entropy loss with a multi-lingual encoder-decoder. In addition, it computes contrastive loss on the representations of the aligned pair (positive example) and randomly selected non-aligned pair (negative example).

Despite these benefits, challenges still remain in multilingual NMT. First, previous work on multilingual NMT does not always perform well as their corresponding bilingual baseline especially on rich resource language pairs (Tan et al., 2019; Zhang et al., 2020; Fan et al., 2020). Such performance gap becomes larger with the increasing number of accommodated languages for multilingual NMT, as model capacity necessarily must be split between many languages (Arivazhagan et al., 2019). In addition, an optimal setting for multilingual NMT should be effective for any language pairs, while most previous work focus on improv-

ing English-centric[1] directions (Johnson et al., 2017; Aharoni et al., 2019; Zhang et al., 2020). A few recent exceptions are Zhang et al. (2020) and Fan et al. (2020), who trained many-to-many systems with introducing more non-English corpora, through data mining or back translation.

In this work, we take a step towards a unified many-to-many multilingual NMT with only English-centric parallel corpora and additional monolingual corpora. Our key insight is to close the representation gap between different languages to encourage transfer learning as much as possible.

As such, many-to-many translations can make the most of the knowledge from all supervised directions and the model can perform well for both English-centric and non-English settings. In this paper, we propose a multilingual COntrastive Learning framework for Translation (mCOLT or mRASP2) to reduce the representation gap of different languages, as shown in Figure 1.

The objective of mRASP2 ensures the model to represent similar sentences across languages in a shared space by training the encoder to minimize the representation distance of similar sentences. In addition, we also boost mRASP2 by leveraging monolingual data to further improve multilingual translation quality. We introduce an effective aligned augmentation technique by extending RAS (Lin et al., 2020) – on both parallel and monolingual corpora to create pseudo-pairs. These pseudo-pairs are combined with multilingual parallel corpora in a unified training framework.

Simple yet effective, mRASP2 achieves consistent translation performance improvements for both English-centric and non-English directions on a wide range of benchmarks. For English-centric directions, mRASP2 outperforms a strong multilingual baseline in 20 translation directions on WMT testsets. On 10 WMT translation benchmarks, mRASP2 even obtains better results than the strong bilingual mBART model. For zero-shot and unsupervised directions, mRASP2 obtains surprisingly strong results on 36 translation directions[2], with 10+ BLEU improvements on average.

[1] "English-centric" means that having English as the source or target language

[2] 6 unsupervised directions + 30 zero-shot directions

2 Methodology

mRASP2 unifies both parallel corpora and monolingual corpora with contrastive learning. This section will explain our proposed mRASP2. The overall framework is illustrated in Figure 1

2.1 Multilingual Transformer

A multilingual neural machine translation model learns a many-to-many mapping function f to translate from one language to another. To distinguish different languages, we add an additional language identification token preceding each sentence, for both source side and target side. The base architecture of mRASP2 is the state-of-the-art Transformer (Vaswani et al., 2017). A little different from previous work, we choose a larger setting with a 12-layer encoder and a 12-layer decoder to increase the model capacity. The model dimension is 1024 on 16 heads. To ease the training of the deep model, we apply Layer Normalization for word embedding and pre-norm residual connection following Wang et al. (2019a) for both encoder and decoder. Therefore, our multilingual NMT baseline is much stronger than that of Transformer big model.

More formally, we define $L = \{L_1, \ldots, L_M\}$ where L is a collection of M languages involving in the training phase. $\mathcal{D}_{i,j}$ denotes a parallel dataset of (L_i, L_j), and $\mathcal{D}$ denotes all parallel datasets. The training loss is cross entropy defined as:

$$\mathcal{L}_{\text{ce}} = \sum_{\mathbf{x}^i, \mathbf{x}^j \in \mathcal{D}} -\log P_\theta(\mathbf{x}^i | \mathbf{x}^j) \qquad (1)$$

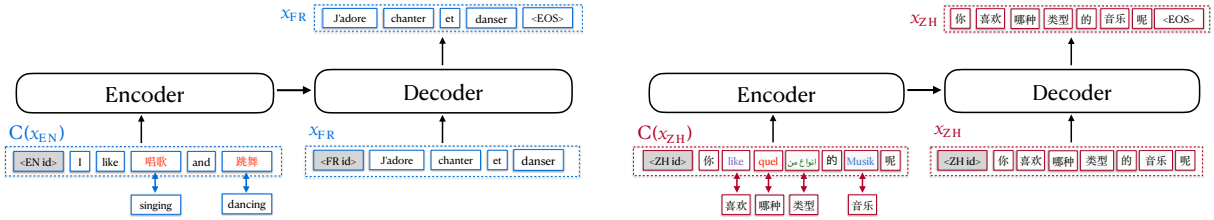
where $\mathbf{x}^i$ represents a sentence in language L_i, and θ is the parameter of multilingual Transformer model.

2.2 Multilingual Contrastive Learning

Multilingual Transformer enables implicitly learning shared representation of different languages. mRASP2 introduces contrastive loss to explicitly bring different languages to map a shared semantic space.

The key idea of contrastive learning is to minimize the representation gap of similar sentences and maximize that of irrelevant sentences. Formally, given a bilingual translation pairs $(\mathbf{x}^i, \mathbf{x}^j) \in \mathcal{D}$, $(\mathbf{x}^i, \mathbf{x}^j)$ is the positive example and we randomly choose a sentence $\mathbf{y}^j$ from language L_j to form a negative example[3] $(\mathbf{x}^i, \mathbf{y}^j)$.

[3] It is possible that $L_j = L_i$



(a) AA for Parallel Corpora

(b) AA for Monolingual Corpora

Figure 2: Aligned augmentation on both parallel and monolingual data by replacing words with the same meaning in synonym dictionaries. It either creates a pseudo-parallel example (left) or a pseudo self-parallel example (right).

The objective of contrastive learning is to minimize the following loss:

$$\mathcal{L}_{\text{ctr}} = -\sum_{\mathbf{x}_i,\mathbf{x}_j \in \mathcal{D}} \log \frac{e^{\text{sim}^+(\mathcal{R}(\mathbf{x}^i),\mathcal{R}(\mathbf{x}^j))/\tau}}{\sum_{\mathbf{y}^j} e^{\text{sim}^-(\mathcal{R}(\mathbf{x}^i),\mathcal{R}(\mathbf{y}^j))/\tau}} \tag{2}$$

where $\text{sim}(\cdot)$ calculates the similarity of different sentences. $+$ and $-$ denotes positive and negative respectively. $\mathcal{R}(s)$ denotes the average-pooled encoded output of an arbitrary sentence s. τ is the temperature, which controls the difficulty of distinguishing between positive and negative examples[4]. In our experiments, it is set to 0.1. The similarity of two sentences is calculated with the cosine similarity of the average-pooled encoded output. To simplify implementation, the negative samples are sampled from the same training batch. Intuitively, by maximizing the softmax term $\text{sim}^+(\mathcal{R}(\mathbf{x}^i), \mathcal{R}(\mathbf{x}^j))$, the contrastive loss forces their semantic representations projected close to each other. In the meantime, the softmax function also minimizes the non-matched pairs $\text{sim}^-(\mathcal{R}(\mathbf{x}^i), \mathcal{R}(\mathbf{y}^j))$.

During the training of mRASP2, the model can be optimized by jointly minimizing the contrastive training loss and translation loss:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \lambda|s|\mathcal{L}_{\text{ctr}} \tag{3}$$

where λ is the coefficient to balance the two training losses. Since $\mathcal{L}_{\text{ctr}}$ is calculated on the sentence-level and $\mathcal{L}_{\text{ce}}$ is calculated on the token-level, therefore $\mathcal{L}_{\text{ctr}}$ should be multiplied by the averaged sequence length $|s|$.

2.3 Aligned Augmentation

We then will introduce how to improve mRASP2 with data augmentation methods, including the introduction of noised bilingual and noised monolingual data for multilingual NMT. The above two types of training samples are illustrated in Figure 2.

Lin et al. (2020) propose Random Aligned Substitution technique (or RAS[5]) that builds code-switched sentence pairs $(C(\mathbf{x}^i), \mathbf{x}^j)$ for multilingual pre-training. In this paper, we extend it to Aligned Augmentation (AA), which can also be applied to monolingual data.

For a bilingual or monolingual sentence pair $(\mathbf{x}^i, \mathbf{x}^j)$[6], AA creates a perturbed sentence $C(\mathbf{x}^i)$ by replacing aligned words from a synonym dictionary[7]. For every word contained in the synonym dictionary, we randomly replace it to one of its synonym with a probability of 90%.

For a bilingual sentence pair $(\mathbf{x}^i, \mathbf{x}^j)$, AA creates a pseudo-parallel training example $(C(\mathbf{x}^i), \mathbf{x}^j)$. For monolingual data, AA takes a sentence $\mathbf{x}^i$ and generates its perturbed $C(\mathbf{x}^i)$ to form a pseudo self-parallel example $(C(\mathbf{x}^i), \mathbf{x}^i)$. $(C(\mathbf{x}^i), \mathbf{x}^j)$ and $(C(\mathbf{x}^i), \mathbf{x}^i)$ is then used in the training by calculating both the translation loss and contrastive loss. For a pseudo self-parallel example $(C(\mathbf{x}^i), \mathbf{x}^i)$, the translation loss is basically the reconstruction loss from the perturbed sentence to the original one.

3 Experiments

This section shows that mRASP2 can achieve substantial improvements over previous many-to-many multilingual translation on a wide range of benchmarks. Especially, it obtains substantial gains on zero-shot directions.

3.1 Settings and Datasets

Parallel Dataset PC32 We use the parallel dataset PC32 provided by Lin et al. (2020). It con-

[4]Higher temperature increases the difficulty to distinguish positive sample from negative ones.

[5]They apply RAS only on parallel data

[6]$\mathbf{x}^i$ is in language L_i and $\mathbf{x}^j$ is in language L_j, where $i, j \in \{L_1, \ldots, L_M\}$

[7]We will release our synonym dictionary

	En-Fr wmt14		En-Tr wmt17		En-Es wmt13		En-Ro wmt16		En-Fi wmt17		Avg	Δ
	→	←	→	←	→	←	→(*)	←	→	←		
bilingual												
Transformer-6(Lin et al., 2020)	43.2	39.8	-	-	-	-	34.3	34.0	-	-	-	
Transformer-12(Liu et al., 2020)	41.4	-	9.5	12.2	33.2	-	34.3	36.8	20.2	21.8	-	
pre-train & fine-tuned												
Adapter (Bapna and Firat, 2019)	-	-	-	-	**35.4**	33.7	-	-	-	-	-	
mBART(Liu et al., 2020)	41.1	-	17.8	22.5	34.0	-	37.7	38.8	22.4	28.5	-	
XLM(Conneau and Lample, 2019)	-	-	-	-	-	-	-	38.5	-	-	-	
MASS(Song et al., 2019)	-	-	-	-	-	-	-	39.1	-	-	-	
mRASP(Lin et al., 2020)	**44.3**	**45.4**	20.0	23.4	-	-	37.6	38.9	**24.0**	28.0	-	
unified multilingual												
Multi-Distillation (Tan et al., 2019)	-	-	-	-	-	-	31.6	35.8	22.0	21.2	-	
m-Transformer	42.0	38.1	18.8	23.1	32.8	33.7	35.9	37.7	20.0	28.2	31.03	
mRASP w/o finetune(**)	43.1	39.2	20.0	25.2	34.0	34.3	37.5	38.8	22.0	29.2	32.33	+1.30
mRASP2	43.5	39.3	**21.4**	**25.8**	34.5	**35.0**	**38.0**	**39.1**	23.4	**30.1**	**33.01**	**+1.98**

Table 1: Performance (tokenized BLEU) on WMT **supervised** translation directions. Consistent BLEU gains are observed in 20 directions (See Appendix) and in this table we pick the representative ones. Different from our work, final BLEU scores of mBART, XLM, MASS and mRASP are obtained by multilingual pre-training and **fine-tuning** on a single direction. Adapter is a trade-off between unified multilingual model and bilingual model (trained on 6 languages on WMT data). Multi-Distillation is improved over Adapter with selective distillation methods. Results for Transformer-6 (6 layers for encoder and decoder) are from Lin et al. (2020). Results for Transformer-12 (12 layers for encoder and decoder separately) are from Liu et al. (2020). (*) Note that for En→Ro direction, we follow the previous setting to calculate BLEU score after removing Romanian dialects. (**) For mRASP w/o finetune we report the results implemented by ourselves, with 12 layers encoder and decoder and our data. Both m-Transformer and our mRASP2 have 12 layers for encoder and decoder.

tains a large public parallel corpora of 32 English-centric language pairs. The total number of sentence pairs is 97.6 million.

We apply AA on PC32 by randomly replacing words in the source side sentences with synonyms from an arbitrary bilingual dictionary provided by (Lample et al., 2018)[8]. For words in the dictionaries, we replace them into one of the synonyms with a probability of 90% and keep them unchanged otherwise. We apply this augmentation in the pre-processing step before training.

Monolingual Dataset MC24 We create a dataset MC24 with monolingual text in 24 languages[9]. It is a subset of the Newscrawl[10] dataset by retaining only those languages in PC32, plus three additional languages that are not in PC32 (Nl, Pl, Pt). In order to balance the volume across different languages, we apply temperature sampling $\tilde{n_i} = \left(n_i / \sum_j n_j\right)^{1/T}$ with T=5 over the dataset, where n_i is the number of sentences in i-th language. Then we apply AA on monolingual data. The total number of sentences in MC24 is 1.01 billion. The detail of data volume is listed in the Appendix.

We apply AA on MC24 by randomly replacing words in the source side sentences with synonyms from a multilingual dictionary. Therefore the source side might contain multiple language tokens (preserving the semantics of the original sentence), and the target is just the original sentence. The replace probability is also set to 90%. We apply this augmentation in the pre-processing step before training. We will release the multilingual dictionary and the script for producing the noised monolingual dataset.

Evaluation Datasets For supervised directions, most of our evaluation datasets are from WMT and IWSLT benchmarks, for pairs that are not available in WMT or IWSLT, we use OPUS-100 instead.

For zero-shot directions, we follow (Zhang et al., 2020) and use their proposed OPUS-100 zero-shot testset. The testset is comprised of 6 languages (Ru, De, Fr, Nl, Ar, Zh), resulting in 15 language pairs and 30 translation directions.

We report de-tokenized BLEU with Sacre-

[8]https://github.com/facebookresearch/MUSE

[9]Bg, Cs, De, El, En, Es, Et, Fi, Fr, Gu, Hi, It, Ja, Kk, Lt, Lv, Ro, Ru, Sr, Tr, Zh, Nl, Pl, Pt

[10]http://data.statmt.org/news-crawl

	En-Nl iwslt2014		En-Pt opus-100		En-Pl wmt20		Nl-Pt -		Avg	Δ
	→	←	→	←	→	←	→	←		
m-Transformer	1.3	7.0	3.7	10.7	0.6	3.2	-	-	4.42	
mRASP	0.7	10.6	3.7	11.6	0.5	5.3	-	-	5.40	+0.98
mRASP2	**10.1**	**28.5**	**18.4**	**30.5**	**6.7**	**17.1**	**9.3**	**8.3**	**18.55**	**+14.13**

Table 2: mRASP2 outperforms m-Transformer in **unsupervised** translation directions by a large margin. We report tokenized BLEU above. For Nl↔Pt, mRASP2 achieves reasonable results after trained only on monolingual data of both sides. The averaged score is calculated without the Nl↔Pt directions.

	Ar		Zh		Nl(*)		
	X→Ar	Ar→X	X→Zh	Zh→X	X→Nl	Nl→X	
Pivot	5.5	17.0	28.5	16.4	2.2	6.0	
m-Transformer	3.7	5.6	6.7	4.1	2.3	**6.3**	
mRASP2	**5.3**	**17.3**	**29.0**	**14.5**	**5.3**	6.1	
	Fr		De		Ru		Avg of all
	X→Fr	Fr→X	X→De	De→X	X→Ru	Ru→X	
Pivot	26.1	22.3	14.4	14.2	16.6	19.9	15.56
m-Transformer	7.7	4.8	4.2	4.8	5.7	4.8	5.05
mRASP2	**23.6**	**21.7**	**12.3**	**15.0**	**16.4**	**19.1**	**15.31**

Table 3: **Zero-Shot:** We report de-tokenized BLEU using sacreBLEU in OPUS-100. We observe consistent BLEU gains in zero-shot directions on different evaluation sets, see Appendix for more details. mRASP2 further improves the quality. We also list BLEU of pivot-based model (X→En then En→Y using m-Transformer) as a reference, mRASP2 only lags behind Pivot by -0.25 BLEU. (*) Note that Dutch(Nl) is not included in PC32.

BLEU (Post, 2018). For tokenized BLEU, we tokenize both reference and hypothesis using Sacremoses[11] toolkit then report BLEU using the `multi-bleu.pl` script[12]. For Chinese (Zh), BLEU score is calculated on character-level.

Experiment Details We use the Transformer model in our experiments, with 12 encoder layers and 12 decoder layers. The embedding size and FFN dimension are set to 1024. We use dropout = 0.1, as well as a learning rate of 3e-4 with polynomial decay scheduling and a warm-up step of 10000. For optimization, we use Adam optimizer (Kingma and Ba, 2015) with ϵ = 1e-6 and β_2 = 0.98. To stabilize training, we set the threshold of gradient norm to be 5.0 and clip all gradients with a larger norm. We set the hyper-parameter $\lambda = 1.0$ in Eq.3 during training. For multilingual vocabulary, we follow the shared BPE (Sennrich et al., 2016) vocabulary of Lin et al. (2020), which includes 59 languages. The vocabulary contains 64808 tokens. After adding 59 language tokens, the total size of vocabulary is 64867.

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

[12]https://github.com/moses-smt/mosesdecoder

4 Experiment Results

This section shows that mRASP2 provides consistent performance gains for supervised and unsupervised English-centric translation directions as well as for non-English directions.

4.1 English-Centric Directions

Supervised Directions As shown in Table 1, mRASP2 clearly improves multilingual baselines by a large margin in 10 translation directions. Previously, multilingual machine translation underperforms bilingual translation in rich-resource scenarios. It is worth noting that our multilingual machine translation baseline is very competitive. It is even on par with the strong mBART bilingual model, which is fine-tuned on a large scale unlabeled monolingual dataset. mRASP2 further improves the performance.

We summarize the key factors for the success training of our baseline[13] m-Transformer: a) The batch size plays a crucial role in the suc-

[13]many-to-many Transformer trained on PC32 as in Johnson et al. (2017) except that we apply language indicator the same way as Fan et al. (2020)

	model	CTL	AA	MC24	Supervised	Unsupervised	Zero-shot
①	m-Transformer				28.65	4.42	5.05
②	mRASP w/o f.t.(*)		✓		29.82	5.40	4.91
③	mRASP2 w/o AA	✓			28.79	4.75	13.55
④	mRASP2 w/o MC24	✓	✓		29.96	5.80	14.60
⑤	mRASP2	✓	✓	✓	**30.36**	**18.55**	**15.31**

Table 4: Summary of average BLEU of mRASP2 w/o AA and mRASP2 in different scenarios. We report averaged tokenized BLEU. For supervised translation, we report the average of 20 directions; for zero-shot translation, we report the average of 30 directions of OPUS-100. mRASP excludes MC24 and contrastive loss from mRASP2. mRASP2 w/o AA only adopts contrastive learning on the basis of m-Transformer. mRASP2 w/o MC24 excludes MC24 from mRASP2. (*) Note that results of mRASP are computed without fine-tuning.

cess of training multilingual NMT. We use 8×4 NVIDIA V100 with update frequency 50 to train the models and each batch contains about 3 million tokens. b) We enlarge the number of layers from 6 to 12 and observe significant improvements for multilingual NMT. By contrast, the gains from increasing the bilingual model size is not that large. mBART also uses 12 encoder and decoder layers. c) We use gradient norm to stable the training. Without this regularization, the large scale training will collapse sometimes.

Unsupervised Directions In Table 2, we observe that mRASP2 achieves reasonable results on unsupervised translation directions. The language pairs of En-Nl, En-Pt, and En-Pl are never observed by m-Transformer. m-Transformer sometimes achieves reasonable BLEU for X→En, e.g. 10.7 for Pt→En, since there are many similar languages in PC32, such as Es and Fr. Not surprisingly, it totally fails on En→X directions. By contrast, mRASP2 obtains +14.13 BLEU score on an average without explicitly introducing supervision signals for these directions.

Furthermore, mRASP2 achieves reasonable BLEU scores on Nl↔Pt directions even though it has only been trained on monolingual data of both sides. This indicates that by simply incorporating monolingual data with parallel data in the unified framework, mRASP2 successfully enables unsupervised translation through its unified multilingual representation.

4.2 Zero-shot Translation for non-English Directions

Zero-shot Translation has been an intriguing topic in multilingual neural machine translation. Previous work shows that the multilingual NMT model can do zero-shot translation directly. However, the translation quality is quite poor compared with pivot-based model.

We evaluate mRASP2 on the OPUS-100 (Zhang et al., 2020) zero-shot test set, which contains 6 languages[14] and 30 translation directions in total. To make the comparison clear, we also report the results of several different baselines. mRASP2 w/o AA only adopt contrastive learning on the basis of m-Transformer. mRASP2 w/o MC24 excludes monolingual data from mRASP2.

The evaluation results are listed in Appendix and we summarize them in Table 3. We find that our mRASP2 significantly outperforms m-Transformer and substantially narrows the gap with pivot-based model. This is in line with our intuition that bridging the representation gap of different languages can improve the zero-shot translation.

The main reason is that contrastive loss, aligned augmentation and additional monolingual data enable a better language-agnostic sentence representation. It is worth noting that, Zhang et al. (2020) achieves BLEU score improvements on zero-shot translations at sacrifice of about 0.5 BLEU score loss on English-centric directions. By contrast, mRASP2 improves zero-shot translation by a large margin without losing performance on English-Centric directions. Therefore, mRASP2 has a great potential to serve many-to-many translations, including both English-centric and non-English directions.

5 Analysis

To understand what contributes to the performance gain, we conduct analytical experiments in this

[14] Arabic, Chinese, Dutch, French, German, Russian

section. First we summarize and analyze the performance of mRASP2 in different scenarios. Second we adopt the sentence representation of mRASP2 to retrieve similar sentences across languages. This is to verify our argument that the improvements come from the universal language representation learned by mRASP2. Finally we visualize the sentence representations, mRASP2 indeed draws the representations closer.

5.1 Ablation Study

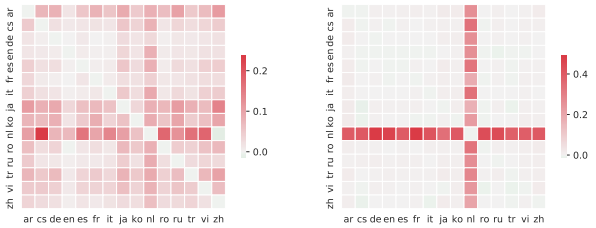
To make a better understanding of the effectiveness of mRASP2, we evaluate models of different settings. We summarize the experiment results in Table 4:

- ①v.s.③: ③ performs comparably with m-Transformer in supervised and unsupervised scenarios, whereas achieves a substantial BLEU improvement for zero-shot translation. This indicates that by introducing contrastive loss, we can improve zero-shot translation quality without harming other directions.
- ②v.s.④: ② performs poorly for zero-shot directions. This means contrastive loss is crucial for the performance in zero-shot directions.
- ⑤: mRASP2 further improves BLEU in all of the three scenarios, especially in unsupervised directions. Therefore it is safe to conjecture that by accomplishing with monolingual data, mRASP2 learns a better representation space.

5.2 Similarity Search

In order to verify whether mRASP2 learns a better representation space, we conduct a set of similarity search experiments. Similarity search is a task to find the nearest neighbor of each sentence in another language according to cosine similarity. We argue that mRASP2 benefits this task in the sense that it bridges the representation gap across languages. Therefore we use the accuracy of similarity search tasks as a quantitative indicator of cross-lingual representation alignment.

We conducted comprehensive experiments to support our argument and experiment on mRASP2 and mRASP2 w/o AA .We divide the experiments into two scenarios: First we evaluate our method



(a) Δacc of mRASP2 w/o AA over m-Transformer

(b) Δacc of mRASP2 over mRASP2 w/o AA

Figure 3: Accuracy Improvements of m-Transformer → mRASP2 w/o AA → mRASP2 for Ted-M. Darker red means larger improvements. mRASP2 w/o AA generally improves accuracy over m-Transformer and mRASP2 especially improves the accuracy X ↔ Nl over mRASP2 w/o AA .

on Tatoeba dataset (Artetxe and Schwenk, 2019), which is English-centric. Then we conduct similar similarity search task on non-English language pairs. Following Tran et al. (2020), we construct a multi-way parallel testset (Ted-M) of 2284 samples by filtering the test split of ted[15] that have translations for all 15 languages[16].

Under both settings, we follow the same strategy: We use the average-pooled encoded output as the sentence representation. For each sentence from the source language, we search the closest sentence in the target set according to cosine similarity.

English-Centric: Tatoeba We display the evaluation results in Table 5. We detect two trends: (i) The overall accuracy follows the rule: m-Transformer < mRASP2 w/o AA < mRASP2. (ii) mRASP2 brings more significant improvements for languages with less data volume in PC32. The two trends mean that mRASP2 increases translation BLEU score in a sense that it bridges the representation gap across languages.

Non-English: Ted-M It will be more convincing to argue that mRASP2 indeed bridges the representation gap if similarity search accuracy increases on zero-shot directions. We list the averaged top-1 accuracy of 210 non-English directions[17] in Table 6. The results show that mRASP2 increases the similarity search accuracy in zero-shot scenario. The results support our argument

[15] http://phontron.com/data/ted_talks.tar.gz

[16] Arabic, Czech, German, English, Spanish, French, Italian, Japanese, Korean, Dutch, Romanian, Russian, Turkish, Vietnamese, Chinese

[17] 15 languages, resulting in 210 directions

Lang	Fr	De	Zh	Ro	Cs	Tr	Ru	NL	PL	Pt
m-Transformer	91.7	96.8	87.0	90.6	84.8	91.1	89.1	25.6	6.3	37.3
mRASP2 w/o AA	91.7	97.3	89.9	91.4	86.1	92.4	90.4	35.7	14.3	46.5
mRASP2	**93.0**	**98.0**	**90.7**	**91.9**	**89.3**	**92.4**	**92.3**	**60.3**	**28.1**	**58.6**

Table 5: **English-Centric:** Sentence retrieval top-1 accuracy on Tatoeba evaluation set. The reported accuracy is the average of En→X and X→En accuracy. mRASP2 outperforms m-Transformer on all directions in English-centric sentence retrieval task.

	Top1 Acc	Δ
m-Transformer	79.8	-
mRASP2 w/o AA	84.4	+4.8
mRASP2	**89.6**	+9.8

Table 6: **Non-English:** The averaged sentence similarity search top-1 accuracy on Ted-M testset. m-Transformer $<$ mRASP2 w/o AA $<$ mRASP2, which is consistent with the results in English-centric scenario.

that our method generally narrows the representation gap across languages.

To better understanding the specifics beyond the averaged accuracy, we plot the accuracy improvements in the heat map in Figure 3. mRASP2 w/o AA brings general improvements over m-Transformer. mRASP2 especially improves on Dutch(Nl). This is because mRASP2 introduces monolingual data of Dutch while mRASP2 w/o AA includes no Dutch data.

5.3 Visualization

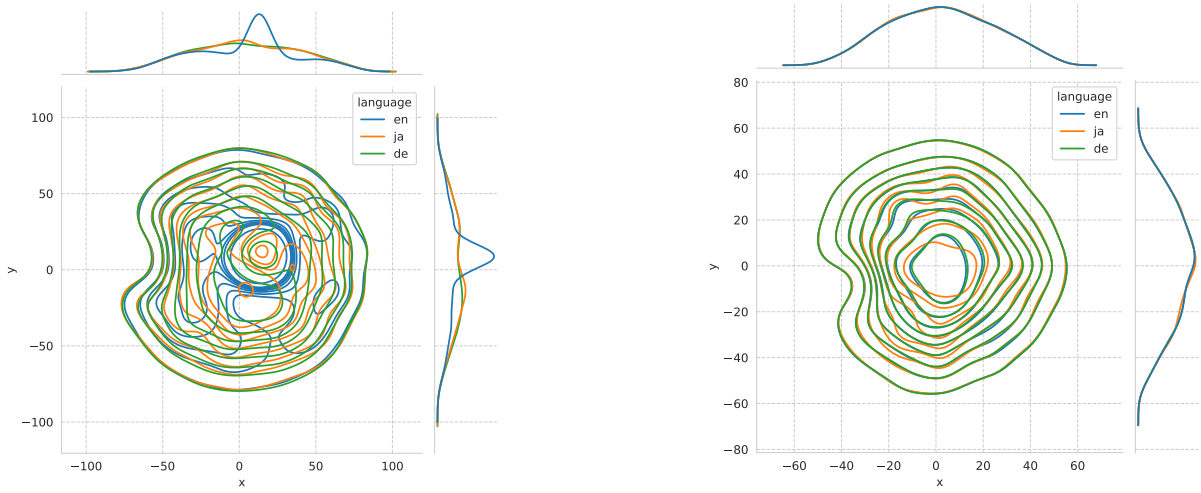
In order to visualize the sentence representations across languages, we retrieve the sentence representation $\mathcal{R}(s)$ for each sentence in Ted-M, resulting in 34260 samples in the high-dimensional space.

To facilitate visualization, we apply T-SNE dimension reduction to reduce the 1024-dim representations to 2-dim. Then we select 3 representative languages: English, German, Japanese and depict the bivariate kernel density estimation based on the 2-dim representations. It is clear in Figure 4 that m-Transformer cannot align the 3 languages. By contrast, mRASP2 draws the representations across 3 languages much closer.

6 Related Work

Multilingual Neural Machine Translation While initial research on NMT starts with building translation systems between two languages, Dong et al. (2015) extends the bilingual NMT to one-to-many translation with sharing encoders across 4 language pairs. Hence, there has been a massive increase in work on MT systems that involve more than two languages (Chen et al., 2018; Choi et al., 2018; Chu and Dabre, 2019; Dabre et al., 2017). Recent efforts mainly focuses on designing language specific components for multilingual NMT to enhance the model performance on rich-resource languages (Bapna and Firat, 2019; Kim et al., 2019; Wang et al., 2019b; Escolano et al., 2020). Another promising thread line is to enlarge the model size with extensive training data to improve the model capability (Arivazhagan et al., 2019; Aharoni et al., 2019; Fan et al., 2020). Different from these approaches, mRASP2 proposes to explicitly close the semantic representation of different languages and make the most of cross lingual transfer.

Zero-shot Machine Translation Typical zero-shot machine translation models rely on a pivot language (e.g. English) to combine the source-pivot and pivot-target translation models (Chen et al., 2017; Ha et al., 2017; Gu et al., 2019; Currey and Heafield, 2019). Johnson et al. (2017) shows that a multilingual NMT system enables zero-shot translation without explicitly introducing pivot methods. Promising, but the performance still lags behind the pivot competitors. Most following up studies focused on data augmentation methods. Zhang et al. (2020) improved the zero-shot translation with online back translation. Ji et al. (2020); Liu et al. (2020) shows that large scale monolingual data can improve the zero-shot translation with unsupervised pre-training. Fan et al. (2020) proposes a simple and effective data mining method to enlarge the training corpus of zero-shot directions. Some work also attempted to explicitly learn shared semantic representation of different languages to im-



(a) m-Transformer

(b) mRASP2

Figure 4: Bivariate kernel density estimation plots of representations after using T-SNE dimensionality reduction to 2 dimension. The blue line is English, the orange line is Japanese and the green line is German. This figure illustrates that the sentence representations are drawn closer after applying mRASP2

prove the zero-shot translation. Lu et al. (2018) suggests that by learning an explicit "interlingual" across languages, multilingual NMT model can significantly improve zero-shot translation quality. Al-Shedivat and Parikh (2019) introduces a consistent agreement-based training method that encourages the model to produce equivalent translations of parallel sentences in auxiliary languages. Different from these efforts, mRASP2 attempts to learn a universal many-to-many model, and bridge the cross-lingual representation with contrastive learning and m-RAS. The performance is very competitive both on zero-shot and supervised directions on large scale experiments.

Contrastive Learning Contrastive Learning has become a rising domain and achieved significant success in various computer vision tasks (Zhuang et al., 2019; Tian et al., 2020; He et al., 2020; Chen et al., 2020; Misra and van der Maaten, 2020). Researchers in the NLP domain have also explored contrastive Learning for sentence representation. Wu et al. (2020) employed multiple sentence-level augmentation strategies to learn a noise-invariant sentence representation. Fang and Xie (2020) applies the back-translation to create augmentations of original sentences. Inspired by these studies, we apply contrastive learning for multilingual NMT.

Cross-lingual Representation Cross-lingual representation learning has been intensively studied in order to improve cross-lingual understanding (XLU) tasks. Multilingual masked language models (MLM), such as mBERT(Devlin et al., 2019) and XLM(Conneau and Lample, 2019), train large Transformer models on multiple languages jointly and have built strong benchmarks on XLU tasks. Most of the previous works on cross-lingual representation learning focus on unsupervised training. For supervised learning, Conneau and Lample (2019) proposes TLM objective that simply concatenates parallel sentences as input. By contrast, mRASP2 leverages the supervision signal by pulling closer the representations of parallel sentences.

7 Conclusion

We demonstrate that contrastive learning can significantly improve zero-shot machine translation directions. Combined with additional unsupervised monolingual data, we achieve substantial improvements on all translation directions of multilingual NMT. We analyze and visualize our method, and find that contrastive learning tends to close the representation gap of different languages. Our results also show the possibilities of training a true many-to-many Multilingual NMT that works well on any translation direction. In future work, we will scale-up the current training to more languages, e.g. PC150. As such, a single model can handle more than 100 languages and outperforms the corresponding bilingual baseline.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3874–3884. Association for Computational Linguistics.

Maruan Al-Shedivat and Ankur P. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1184–1197. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguistics*, 7:597–610.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1538–1548. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Yun Chen, Yang Liu, Yong Cheng, and Victor O. K. Li. 2017. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1925–1935. Association for Computational Linguistics.

Yun Chen, Yang Liu, and Victor O. K. Li. 2018. Zero-resource neural machine translation with multi-agent communication game. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5086–5093. AAAI Press.

Gyu-Hyeon Choi, Jong-Hun Shin, and Young Kil Kim. 2018. Improving a multi-source neural machine translation model with corpus extension for low-resource languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Chenhui Chu and Raj Dabre. 2019. Multilingual multi-domain adaptation approaches for neural machine translation. *CoRR*, abs/1906.07978.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 7057–7067.

Anna Currey and Kenneth Heafield. 2019. Zero-resource neural machine translation with monolingual pivot data. In *Proceedings of the 3rd Workshop on Neural Generation and Translation@EMNLP-IJCNLP 2019, Hong Kong, November 4, 2019*, pages 99–107. Association for Computational Linguistics.

Raj Dabre, Fabien Cromierès, and Sadao Kurohashi. 2017. Enabling multi-source neural machine translation by concatenating source sentences in multiple languages. *CoRR*, abs/1702.06135.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1723–1732. The Association for Computer Linguistics.

Carlos Escolano, Marta R. Costa-jussà, José A. R. Fonollosa, and Mikel Artetxe. 2020. Training multilingual machine translation by alternately freezing language-specific encoders-decoders. *CoRR*, abs/2006.01594.

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, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125.

Hongchao Fang and Pengtao Xie. 2020. CERT: contrastive self-supervised learning for language understanding. *CoRR*, abs/2005.12766.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1258–1268. Association for Computational Linguistics.

Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2017. Effective strategies in zero-shot neural machine translation. *CoRR*, abs/1711.07893.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9726–9735. IEEE.

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. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 115–122. AAAI Press.

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.

Yunsu Kim, Yingbo Gao, and Hermann Ney. 2019. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1246–1257. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

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.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

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.

Ishan Misra and Laurens van der Maaten. 2020. Self-supervised learning of pretext-invariant representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 6706–6716. IEEE.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. 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 ACL (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Xu Chen, Sneha Reddy Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2827–2835. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on*

Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI*, volume 12356 of *Lecture Notes in Computer Science*, pages 776–794. Springer.

Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual retrieval for iterative self-supervised training. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. 2019a. Learning deep transformer models for machine translation. In *Proceedings of the 57th ACL*, pages 1810–1822, Florence, Italy. Association for Computational Linguistics.

Yining Wang, Long Zhou, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2019b. A compact and language-sensitive multilingual translation method. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1213–1223. Association for Computational Linguistics.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. CLEAR: contrastive learning for sentence representation. *CoRR*, abs/2012.15466.

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 ACL*, pages 1628–1639, Online. Association for Computational Linguistics.

Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. 2019. Local aggregation for unsupervised learning of visual embeddings. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6001–6011. IEEE.

A Case Study

We plot the location of multi-way parallel sentences in the representation space of mRASP2 in Figure 5 and list sentences number 1 and 100 in Table 7

B Details of Evaluation Results

We list detailed results of evaluation on a wide range of test sets.

B.1 Results on OPUS-100

Detailed results on OPUS-100 zero-shot evaluation set are listed in Table 8

B.2 Results on WMT

Detailed results on WMT evaluation set are listed in Table 9

C Example of AA

We show two results of sentences after AA in Figure 6

D Details of MC24

We describe the detail of MC24 in Table 10

Id	Language	Sentence
1	De	Was sie alle eint, ist, dass sie sterben werden.
	En	The one thing that all of them have in common is that they're going to die.
	Ja	１つ全員に共通して言えるのは皆いずれ死ぬということです
100	De	Rechts seht Ihr meinen Kollegen Sören , der sich wirklich in dem Raum befindet.
	En	On the right side you can see my colleague Soren , who 's actually in the space.
	Ja	右側には同僚・ソーレンが見えます実際その場所にいたのです

Table 7: Case Study: Parallel sentences distributed in English, German and Japanese.

m-Transformer

	Ar	Zh	Nl	Fr	De	Ru	Avg
Ar	-	9.2	1.2	7.6	1.8	8.2	5.6
Zh	4.7	-	0.8	7.7	1.7	5.8	4.1
Nl	1.9	5.1	-	10.8	9.9	3.7	6.3
Fr	3.9	6.5	3.7	-	4.3	5.3	4.8
De	3.1	4.4	4.5	6.5	-	5.5	4.8
Ru	4.8	8.4	1.5	5.9	3.2	-	4.8
Avg	3.7	6.7	2.3	7.7	4.2	5.7	**5.05**

mRASP2 w/o AA

	Ar	Zh	Nl	Fr	De	Ru	Avg
Ar	-	26.1	1.2	19.1	10.5	12.5	13.9
Zh	5.6	-	0.9	32.1	8.0	17.3	12.8
Nl	2.3	5.5	-	10.3	10.3	3.8	5.6
Fr	5.6	41.5	3.7	-	18.0	19.5	18.8
De	4.6	19.9	4.4	23.0	-	13.6	13.1
Ru	5.9	37.4	1.5	30.1	12.2	-	17.4
Avg	4.8	26.1	2.3	22.9	11.8	13.3	**13.55**

mRASP

	Ar	Zh	Nl	Fr	De	Ru	Avg
Ar	-	5.7	1.6	1.2	6.8	4.0	3.9
Zh	4.0	-	4.2	3.8	5.4	2.9	4.1
Nl	3.0	7.5	-	4.4	7.8	2.6	5.1
Fr	2.8	14.8	13.3	-	5.4	7.4	6.4
De	5.3	6.1	2.5	1.4	-	3.4	3.7
Ru	5.2	6.7	1.5	1.0	5.6	-	4.0
Avg	4.1	8.2	4.6	2.4	6.2	4.1	**4.91**

mRASP2 w/o MC24

	Ar	Zh	Nl	Fr	De	Ru	Avg
Ar	-	28.8	1.0	20.9	7.9	15.6	14.8
Zh	6.3	-	0.7	33.8	5.9	20.0	13.3
Nl	3.2	8.1	-	16.3	14.3	6.0	9.6
Fr	6.6	41.5	3.7	-	16.7	21.4	19.1
De	6.1	21.3	4.6	24.3	-	15.0	14.3
Ru	7.1	38.0	1.1	30.6	11.1	-	17.6
Avg	5.9	27.5	2.2	25.2	11.2	15.6	**14.60**

mRASP2

	Ar	Zh	Nl	Fr	De	Ru	Avg
Ar	-	32.5	3.2	22.8	11.2	16.7	17.3
Zh	6.5	-	1.9	32.9	7.6	23.7	14.5
Nl	1.7	8.2	-	7.5	10.2	2.9	6.1
Fr	6.2	42.3	7.5	-	18.9	24.4	21.7
De	4.9	21.6	9.2	24.7	-	14.4	15.0
Ru	7.1	40.6	4.5	29.9	13.5	-	19.1
Avg	5.3	29.0	5.3	23.6	12.3	16.4	**15.31**

Pivot

	Ar	Zh	Nl	Fr	De	Ru	Avg
Ar	-	31.4	1.0	22.9	13.5	16.4	17.0
Zh	7.3	-	0.8	37.7	11.9	24.2	16.4
Nl	1.7	4.9	-	10.1	9.7	3.7	6.0
Fr	6.8	44.1	3.6	-	21.4	23.2	22.3
De	4.9	20.8	4.3	25.3	-	15.5	14.2
Ru	6.7	41.5	1.4	34.5	15.5	-	19.9
Avg	5.5	28.5	2.2	26.1	14.4	16.6	**15.56**

Table 8: Detailed de-tokenized BLEU on OPUS-100 zero-shot test set. Note that results of mRASP are computed without fine-tuning.

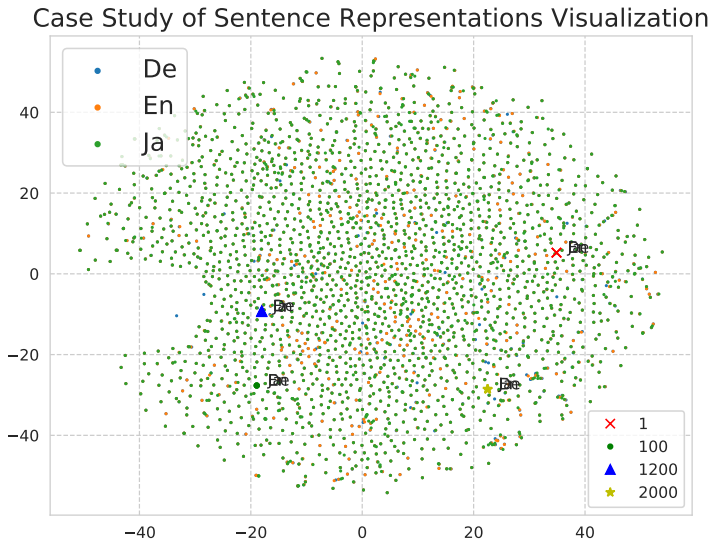


Figure 5: Case Study: Examples of representations of multi-way parallel sentences on mRASP2 representation space. We can observe that similar sentences overlap perfectly on the space. Numbers in the legend means the id of sentence in Ted-M (See Table 7 for detailed sentences). We can clearly observe that similar sentences are clustered to the neighboring location.

1	Original (En)	One more point is lost in this debate: that the EU is proposing far fewer rules now.
	AA	One высокого πόντος той perduti العام tento diskusijos : tuo cette EU is soovitab 遠く 低い регламент घटे .
2	Original (En)	" If we don 't win , there will be some inquiries of why we haven't , " Graves told BBC Radio Leeds.
	AA	" If noi annetada 't ויטוריה , そ こ хочу jet sometime αιτήσεις seine kuna bize haven't , " Graves erzählte BBC Radio Leeds.

Figure 6: Two examples of sentences with its noised version after AA

	En-Fr wmt14		En-De wmt14		En-Zh wmt17		En-Ro wmt16		En-Cs wmt16			
	→	←	→	←	→	←	→	←	→	←		
m-Transformer	42.0	38.1	27.1	34.2	32.8	24.2	26.9	37.7	20.9	31.3		
mRASP2 w/o AA	42.1	38.7	26.8	34.6	33.2	24.7	26.6	37.5	20.8	31.5		
mRASP	43.1	39.2	29.2	34.6	34.8	24.8	28.2	38.8	22.5	32.1		
mRASP2 w/o MC24	43.3	39.3	29.1	34.7	**35.0**	**24.5**	28.4	39.0	22.4	32.5		
mRASP2	**43.5**	**39.3**	**29.7**	**35.0**	34.6	23.8	**28.7**	**39.1**	**24.3**	**33.1**		
	En-Tr wmt16		En-Ru wmt19		En-Fi wmt17		En-Es wmt13		En-It wmt09		Avg	Δ
	→	←	→	←	→	←	→	←	→	←		
m-Transformer	18.2	24.3	17.0	22.6	20.0	28.2	32.8	33.7	29.0	32.0	28.65	
mRASP2 w/o AA	18.2	24.8	17.6	23.2	20.0	27.8	33.1	33.2	29.2	32.2	28.79	+0.14
mRASP	20.0	25.2	18.6	23.3	22.0	29.2	34.0	34.3	30.1	32.4	29.82	+1.17
mRASP2 w/o MC24	20.4	25.7	18.6	**23.4**	22.0	29.4	34.1	34.3	30.4	32.6	29.96	+1.31
mRASP2	**21.4**	**25.8**	**19.2**	23.2	**23.4**	**30.1**	**34.5**	**35.0**	**30.8**	**32.6**	**30.36**	**+1.71**

Table 9: Tokenized BLEU score on public WMT testsets. mRASP2 w/o AA only adopt contrastive learning on the basis of m-Transformer. mRASP excludes MC24 and contrastive loss from mRASP2. mRASP2 w/o MC24 excludes monolingual data from mRASP2. Note that results of mRASP are computed without fine-tuning.

Lanuage	Original Num.	Sampling Ratio	% of replaced tokens	Final Num.
bg	37870628	1.58	/	59839631
cs	75808960	0.89	0.29	67118121
de	319938740	0.29	0.40	91985353
el	4178943	5.50	0.35	22980970
en	224446700	0.38	0.62	85785847
es	17632409	1.24	0.60	21783966
et	4978345	7.82	0.28	38925275
fi	19954908	2.57	0.29	51368970
fr	85274195	0.84	0.54	71760116
gu	530747	35.26	/	18716499
hi	6240797	1.85	0.46	11521321
it	39170950	1.56	0.47	61064797
ja	3250665	11.14	0.15	36225302
kk	1853728	18.30	/	33926819
lt	2446627	13.02	0.16	31857781
lv	10942229	4.30	0.35	47032289
ro	20094801	2.62	0.34	52685562
ru	89373208	0.79	0.29	70839964
sr	3801560	10.30	/	39167541
tr	16337598	3.03	0.29	49502982
zh	4238918	8.66	0.15	36706289
nl	1177713	1.00	0.52	1177713
pl	3404714	1.00	?	3404714
pt	9103090	1.00	?	9103090
SUM				1014480912

Table 10: Detail of MC24, '?' means the data is missing, and '/' means the corresponding language is not contained in the synonym dictionary.