ISSN:2790-1688 Volume-14-(2025)

MCL-MT: Multi-Level Contrastive Learning for Many-to-many Multilingual Neural Machine Translation

Haolun Ran

Anhui University Hefei, China

Abstract. Currently, mainstream multilingual translation models are trained mainly on English- related language pairs. These systems usually perform well on the English-related directions, known as supervised directions, while the translation performance on non-English directions (zero-resource directions) is weak. Therefore, a method called mRASP2 has been proposed which integrates monolingual corpus and bilingual corpus under a unified training framework through contrastive learning and alignment enhancement methods, so that it can make full use of the corpus, learn bet- ter language-independent representations, and thus improve the performance of multilingual translation. In this paper, we propose a method to train an NMT model. Unlike the sentence- level alignment used in most previous stud- ies, this paper uses MCL-MT to explicitly in- tegrate word-level information of each pair of parallel sentences into contrastive learning. English-centric translation directions show su- perior performance with MCL-MT's method- ology in comparison to the pretrained and fine-tuned model mBART. On non-English translations, MCL-MT offers an average 10+ BLEU score increase when compared to the multilingual Transformer baseline.

1. Introduction

As globalization accelerates, people cannot do without cross-language communication in activi- ties such as diplomacy, foreign trade, and tourism. However, traditional human translation has limita- tions such as its high cost and poor real-time per- formance, so it is only applicable to a small num- ber of scenarios. The emergence of machine trans- lation has broken these restrictions and greatly expanded its application scenarios. Modern ma- chine translation models translate the input sen- tences into sentences of another language through

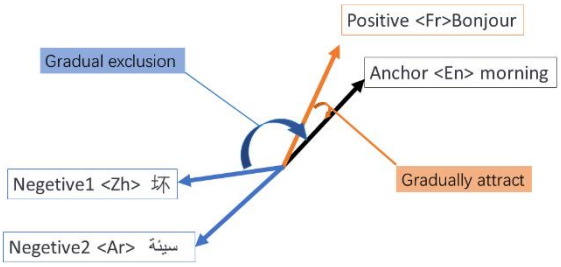

Figure 1: The previous mRASP2 method calculates the standard cross-entropy loss by defining a cross-entropy loss function through a multilingual encoder-decoder. where the contrast loss is calculated by aligning the embedding of pairs of positive and negative examples.

neural network models. The traditional approach to solving machine translation problems between two languages often involves learning the features of each language separately and then attempting to match them, but this neglects the significant differences in feature expression between the two languages, resulting in poorer model performance. To improve the translation level between various languages, especially non-English languages, it is increasingly important to effectively utilize the features of different languages to build models. At present, the dominant approach in neural machine translation relies on the "encoder-decoder" framework. In this structure, the encoder takes the source language sentence and transforms it into a continuous space vector, while the decoder generates the target language sentence based on this vector.

Traditional machine translation can only sup- port single-direction translation, while multilingual machine translation models have the abil- ity to support multiple translation directions at the same time. Multilingual machine transla- tion has attracted widespread attention from researchers and engineers in recent years due to its low deployment cost, transfer learning effect, and other advantages. As a pre-work of mRASP2, mRASP mainly proposed the idea of "machine translation pre-training" to efficiently use differ- ent language pairs corpus. mRASP introduces word-level alignment information based on par- allel dictionaries and replaces it with alignment substitutions (RAS). Experiments have shown that RAS does indeed close the distance between syn - onyms in high-dimensional representations and in- directly closes the distance between synonymous sentences. mRASP2 introduces contrastive learn- ing to further close the distance between synony- mous sentences and explicitly closes the distance between synonymous sentences. The present study introduces a multi-level contrastive learning (MCL-MT) framework to further enhance the cross-lingual capabilities of the pre-trained model. In this approach, translation parallel data is utilized to encourage the model to generate similar semantic embeddings for different languages. Distinguishing itself from prior research that primarily focuses on sentence-level alignment, this paper explicitly incorporates word-level information from each pair of parallel sentences in the contrastive learning process. Additionally, to mitigate the impact of floating point errors during mini-batch training, the Cross-Zero-Noise Contrastive Estimation loss is adopted. In practical cases, the amount of monolin- gual corpus is much larger than the amount of parallel corpus. For traditional unidirectional ma- chine translation models, single-language corpus can be used to enhance translation performance through back-translation technology. For multi- lingual translation, although the back-translation technology is still effective, the process is too long and too complicated. Now, MCL-MT trains monolingual and bilingual data under a unified framework in order to fully and simply utilize the widely available various corpora.

2. Methodology

Through the improved mRASP2 algorithm, the contrastive learning will be extended to the word level. This section will explain how we pro- posed MCL-MT. Besides, The overall framework is roughly illustrated in Figure 3.

2.1 Multilingual Transformer

We combine pseudo-pairs with multilingual par- allel texts in a unified training environment. For each sentence, we add an additional language identification token before it for both source and target language. We use a Transformer model with 12-layer encoders and decoders in a larger setting to increase the capacity of our model, whose di- mension is 1024 on 16 heads. To simplify the training process, In our approach, we utilize Layer Normalization for word embedding and apply pre-norm residual connections in both the encoder and decoder components. This design choice makes our multilingual Neural Machine Translation (NMT) baseline significantly more robust and capable compared to the Transformer big model.

Let $S = \{S1, \ldots, SM\}$ represent a collection of M languages involved in the training phase. $D_{i,j}$ denotes a parallel dataset consisting of sentences from language Si and its translation in language Sj. The set D represents all parallel datasets used in training.

Furthermore, z denotes the i-th word in the k-th language. The training loss is defined using cross entropy, which measures the dissimilarity between the predicted probability distribution and the true distribution:

$$\mathcal{L}_{\text{ce}} = \sum_{k=1}^{M} \sum_{z_k^i, z_k^j \in \mathcal{D}} -\log P_\theta\left(z_k^i | z_k^j\right)$$

where θ in the above equation represents the pa- rameter of multilingual Transformer model.

2.2 Multi-language Contrastive Learning

A multilingual translator maps similar words be- tween different languages into the same simi- lar space through the introduction of contrastive learning. The original mRASP2 method only achieved contrastive learning at the phrase level, resulting in less precise results in multilingual problems. Now a new method is used to define a new contrastive learning loss at this level.

The critical thought in contrastive learning is minimizing the gaps in the representation of similar clauses and maximizing the gaps in the representation of unrelated clauses. For normally, a specific pair of bilingual translators is given $(zi, zj) \in D$, (zi, zj) is the positive example and we select a clause at random yj from language Lj which is used to form a negative example1 (zi, zj). The goal of contrastive learning is to minimize the loss :

$$\mathcal{L}_{\text{fcl}} = -\sum_{k^i, k^j \in M} \sum_{z_i, z_j \in \mathcal{D}} \log \frac{e^{\text{sim}^+\left(\mathcal{R}(z_k^i), \mathcal{R}\left(z_k^j\right)\right)/\tau}}{\sum_{z^j} e^{\text{sim}^-\left(\mathcal{R}\left(z_{k^i}^i\right), \mathcal{R}\left(z_{k^j}^j\right)\right)/\tau}}$$

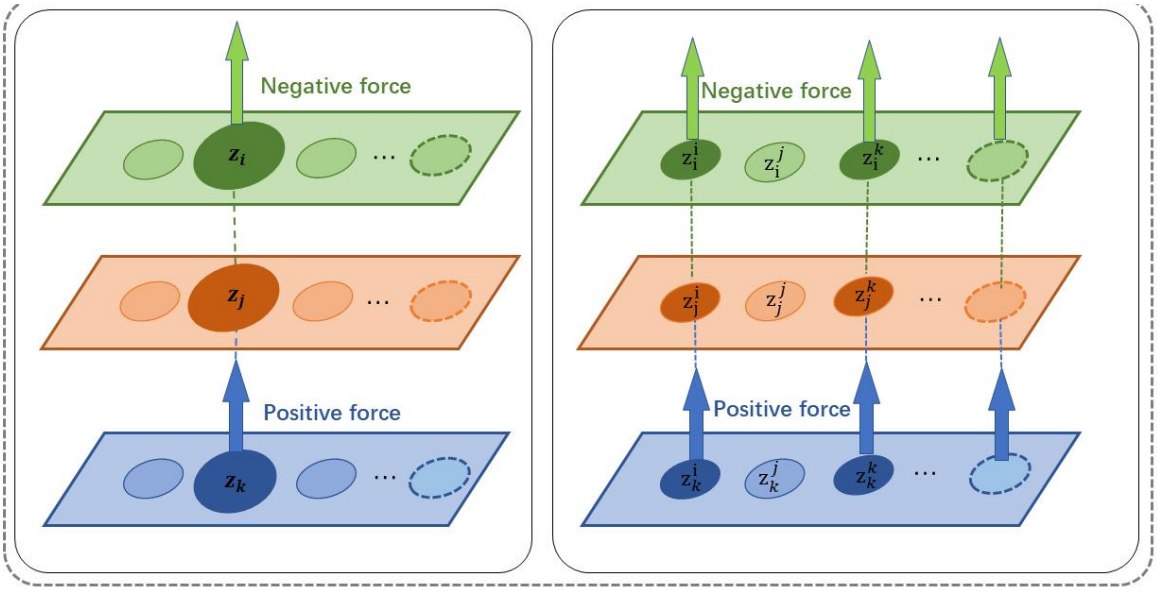

Figure 2: In contrast to the sentence-level alignment used in the m-RASP2 study, we used MCL-MT comparative learning to pinpoint the word level of each pair of parallel sentences

In our experiments, the contrastive loss is set to a value of 0.1. The similarity between two sentences is calculated using the cosine similarity of their average-pooled encoded outputs. To simplify the implementation, negative samples are sampled from the same training batch.Intuitively, by maximizing the softmax term$\text{sim}^+\left(\mathcal{R}\left(z_k^i\right), \mathcal{R}\left(z_k^j\right)\right)$, the contrastive loss encourages the semantic representations of positive pairs to be projected close to each other. At the same time, the softmax function also minimizes the semantic similarity between non-matched pairs, represented by$\text{sim}^-\left(\mathcal{R}\left(z_k^i\right), \mathcal{R}\left(z_k^j\right)\right)$. This helps create a clear distinction between matching and non-matching pairs.

At the same time, we take the sentence-level loss function into account, as shown below:

$$\mathcal{L}_{\text{fsl}} = -\sum_{z_i, z_j \in \mathcal{D}} \log \frac{e^{\text{sim}^+\left(\mathcal{R}\left(z^i\right), \mathcal{R}\left(z^j\right)\right)/\tau}}{\sum_{y^j} e^{\text{sim}^-\left(\mathcal{R}\left(z^i\right), \mathcal{R}\left(y^j\right)\right)/\tau}} \quad (3)$$

While training process of MCL-MT, the model is optimized by simultaneously minimizing the contrastive training loss and the translation loss:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_1 |s| \mathcal{L}_{fsl} + \lambda_2 |s| \mathcal{L}_{fcl} \quad (4)$$

where λ1 and λ2 are the coefficient to balance the two training losses. Since Lce is calculated on the sentence-level, Lfclis calculated on the token-level and L_{fsl} is calculated on the word-level. We have divided it into three levels. Therefore, L_{fsl} should be multiplied by the averaged sequence length |s| .

2.3 Aligned Augmentation

In order to improve MCL-MT, Figure 2 depict how data augmentation can be done with the introduc- tion of two types of noisy training samples: bilin- gual and monolingual data for multilingual NMT.

Random Aligned has been proposed by Lin et al. (2020) substitution technique (or RAS3) that builds code-switched sentence pairs (xi, xj) for multilingual pre-training. With this paper, we introduce Aligned Augmentation (AA) to the lit- erature, which can be utilized to improve the per- formance of monolingual datasets. In the case of a bilingual or monolingual sentence pair (z_i, z_j), a data augmentation technique called AA (Adversarial Autoencoder) is employed. AA introduces perturbed sentences by replacing aligned words with their synonyms from a predefined synonym dictionary. Specifically, for each word present in the synonym dictionary, there is a 90% probability of randomly replacing it with one of its synonyms. For bilingual sentence pairs (z_i, z_j), AA generates pseudo-parallel training examples in the form of $(C(z_i), z_j)$, where $C(z_i)$ represents the perturbed version of zi. In the case of monolingual data, AA takes a sentence zi and generates its perturbed counterpart $C(z_i)$, forming a pseudo self-parallel example $(C(z_i), z_i)$. These pseudo-parallel examples, namely $(C(z_i), z_j)$ and $(C(z_i), z_i)$, are then used in training by calculating both the translation loss and contrastive loss. For a pseudo self-parallel example $(C(z_i), z_i)$, the translation loss is basically the reconstruction loss from the perturbed sentence to the original one.

3. Experiments

Now we reveal that MCL-MT is able to achieve significant enhancements compared to previous many-to-many multilingual translations across a variety of benchmarks.Especially, it obtains sub- stantial gains on zero-shot directions.

3.1 Settings and Datasets

Parallel Dataset PC32 In this paper, we use the parallel dataset PC32 provided by Lin et al. (2020). PC32 is an open-source dataset that pro- vides high-quality. PC32 contains 97.6 million pairs of sentences in 32 English-centric language pairs. It is a large public parallel corpus with a wide range of languages.

Before training, we will use the augmentation method of AA on PC32 by randomly changing words of the source side sentences to synonyms from a bilingual dictionary. For words found in the dictionaries, they will be replaced into one of the synonyms with a probability of 90% while the remaining words are kept unchanged.

In order to enhance the diversity and robustness of the MC24 dataset, we incorporate the AA (Adversarial Autoencoder) technique. Specifically, we randomly replace words in the source side sentences with synonyms obtained from a multilingual dictionary. This augmentation process introduces the possibility of multiple language tokens within the source side, while preserving the semantic meaning of the original sentence. Meanwhile, the target side remains unchanged and consists of the original sentence. The probability of word replacement is set to 90%.

Evaluation Datasets Regarding supervised directions, the majority of our evaluation datasets are sourced from WMT benchmarks. In cases where pairs of data are not accessible through WMT or IWSLT, we make use of OPUS-100.

In order to evaluate zero-shot translation capabilities, we utilize the OPUS-100 zero-shot testset. This testset consists of 6 languages (Russian, German, French, Dutch, Arabic, Chinese).

To measure translation quality, we report de-tokenized BLEU scores using the SacreBLEU metric (Post, 2018). For tokenized BLEU scores, both the reference and hypothesis sentences are tokenized using the Sacremoses toolkit. then report

Experiment Details In our experimental setup, we utilize the Transformer model, which consists of 12 encoder layers and 12 decoder layers. The embedding size and the dimension of the feed-forward network (FFN) are both set to 1024. To prevent overfitting, we apply a dropout rate of 0.1.

In terms of optimization, we employ the Adam optimizer (Kingma and Ba, 2015) with a small epsilon value (ϵ) of 1e-6 and a second moment decay rate ($\beta 2$) of 0.98. We use a learning rate of 3e-4, combined with polynomial decay scheduling and a warm-up step of 10000. To ensure stable training, we set the threshold for the gradient norm to be 5.0 and clip all gradients with a larger norm.

During training, we incorporate the hyper-parameter $\lambda = 1.0$ in Eq.4, which plays a role in the loss function.

4. Experiment Results

The upcoming section suggests that MCL-MT has enhanced precision in both supervised and unsu- pervised translations that involved not only En- glish but also other specific languages.

4.1 English-Centric Directions

MCL-MT markedly improved the multilingual Baseline in the 10 translation directions, as shown in Table 1, In the past, multilingual machine translation was not as successful as bilin- gual translation in scenarios with abundant re- sources. It is important to note that our baseline for multilingual machine translation is highly compet- itive. Our model has improved the multilingual baseline in multiple directions. The final result is even comparable to the current mainstream translation model mBERT.

We summarize the main factors that contribute to the successful training.

1) Batch size plays a critical factor in the successful training.

2) We expanded the layers up from 6 to 12 and noticed a remarkable improvement for the

	En-Fr wmt14		En-Tr wmt17		En-Es wmt13		En-Ro wmt16		En-Fi wmt17		Avg	Δ
	→	←	→	←	→	←	→(*)	←	→	←		
bilingual												
Transformer	43.2	39.8	-	-	-	-	34.3	34.0	-	-	-	
Transformer-12	41.4	-	9.5	12.2	33.2	-	34.3	36.8	20.2	21.8	-	
Pre trained and fine tuned												
Adapter	-	-	-	-	**35.4**	33.7	-	-	-	-	-	
MASS	41.1	-	17.8	22.5	34.0	-	37.7	38.8	22.4	28.5	-	
XLM	-	-	-	-	-	-	-	38.5	-	-	-	
mBART	-	-	-	-	-	-	-	39.1	-	-	-	
mRASP	**44.3**	**45.4**	20.0	23.4	-	-	37.6	38.9	**24.0**	28.0	-	
unified multilingual												
Multi-Distillation	-	-	-	-	-	-	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	
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**
MCL-MT	43.9	39.4	**22.4**	**26.2**	34.2	**35.7**	**38.5**	**39.4**	22.0	29.2	32.33	+1.30

Table 1: In WMT, we noticed a uniform BLEU advantage in 20 directions regarding the performance of the supervised translation strategy (tokenized BLEU). In this table, we picked representative gains. Distinct from our work, mBART, XLM, MASS, and mRASP's final BLEU scores were acquired through multilingual pre-training and single-direction fine-tuning. Adapters

are a balance between unified multi-lingual models and bilingual models (trained with 6 languages on WMT data). Multi-distillation was refined by selective distillation techniques. The output of Transformer-6 (six layers for encoding and decoding) was provided by Lin et al (2020). Liu et al(2020) supplied the result of Transformer-12 (12 layers for encoding and decoding)Please take note that for the En- Ro direction, the BLEU scores were calculated after erasing the Romanian dialects by following the preceding settings. For the non-fine-tuned mRASP, we reported the findings from our own implementation, containing 12- layer encoders and decoders, as well as our data. Both m-Transformer and our MCL-MT have 12 layers each, used for encoding and decoding.

	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	
mRASP2	0.7	10.6	3.7	11.6	0.5	5.3	-	-	5.40	+0.98
MCL-MT	**10.1**	**28.5**	**18.4**	**30.5**	**6.7**	**17.1**	**9.3**	**8.3**	**18.55**	**+14.13**

Table 2: MCL-MT significantly outperforms m-Transformer in unsupervised translation directions. The averaged score is calculated without the Nl↔Pt directions.

	En		Zh		Ge(*)		
	X→En	En→X	X→Zh	Zh→X	X→Ge	Ge→X	
Pivot	5.6	16.9	27.4	15.8	2.1	5.9	
m-Transformer	3.7	5.6	6.7	4.1	2.3	**6.3**	
MCL-MT	**5.4**	**17.4**	**29.2**	**14.7**	**5.5**	6.0	
	Fr		De		Ru		Avg of all
	X→Sp	Fr→Sp	X→Da	Da→X	X→Ru	Ru→X	
Pivot	25.9	21.2	13.8	13.1	17.2	20.0	18.53
m-Transformer	8.3	5.2	4.5	5.6	5.6	3.6	5.47
MCL-MT	**23.8**	**22.1**	**12.4**	**15.3**	**16.5**	**19.4**	**15.4**

Table 3: **Zero-Shot:** We observed consistent BLEU improvement in zero-shot directions when we used sacre- BLEU to de-tokenize BLEU on the OPUS-100 dataset. The results, detailed in the Appendix, demonstrate that MCL-MT further enhances the performance.

	model	ce	fsl	fcl	Supervised	Unsupervised	Zero-shot
①	mRASP2				27.65	4.42	5.05
②	MCL-MT w/o ce		✓	✓	28.82	5.40	4.91
③	MCL-MT w/o fsl	✓		✓	27.79	4.75	13.55
④	MCL-MT w/o fcl	✓	✓		28.96	5.80	14.60
⑤	MCL-MT	✓	✓	✓	**29.27**	**17.45**	**15.72**

Table 4 presents the average BLEU scores. The supervised translation results are averaged over 20 translation directions, while the zero-shot translation results are averaged over 20 translation directions.

multilingual NMT. By contrast, the gains from in- creasing the bilingual model size is not that large. mBART also uses 12 encoder and decoder lay- ers. c) We use gradient norm to stable the training. Without this regularization, the large scale training will collapse sometimes.

Unsupervised Directions As seen in Table 2

MCL-MT achieves satisfactory results on unsu-pervised translation directions, even for those language pairs that had never been seen before by mRASP2 such as En-Nl, En-Pt,andEn-Pl. Due to the presence of numerous similar languages in PC32, such as Es (Spanish)

and Fr (French), mRASP2 occasionally manages to attain reasonable BLEU scores for X→En translation directions. For instance, it achieves a BLEU score of 10.7 for Pt (Portuguese)→En (English). However, it is not surprising that mRASP2 fails completely when translating from En to X directions. Meanwhile, MCL-MT achieve averages a +14.13 BLEU score on these directions all that without the introduction of any explicit supervision signals.

Moreover, MCL-MT gains reasonable BLEU scores on Nl↔Pt translations even though it was only trained on monolingual data from each side. This implies that the integration of an all-encompassing structure for monolingual and par- allel data within a single system is satisfactory for effective unsupervised translations.

4.2 Zero-shot Translation for non-English Directions

Research into zero-shot translation has been an interesting focus in the field of multilingual neu- ral machine translation. While past studies have demonstrated that a multilingual NMT model can translate without any prior preparation, the re- sultant translations are comparatively less precise than pivoting-based models.

We examined the performance of MCL-MT on the OPUS- 100 (Zhang et al., 2020) zero-shot test

set, which consists of 6 languages9 and 30 trans- lations within them. We also compared our results with a variety of other baselines, to ensure an ac- curate comparison. MCL-MT w/o fcl only per- formed translations on sentence level.

The assessment results can be found in the appendix, while a summary of these results is provided in Table 3. Our findings indicate that mRASP2 significantly outperforms m-Transformer and considerably reduces the disparity with the pivot-based model. This outcome aligns with our intuition that addressing the representation gap between different languages can enhance zero-shot translation performance. The underlying cause of this im- provement is due to contrastive loss, aligned aug- mentation, and additional monolingual data, al- lowing for a more generalizable sentence repre- sentation across languages.

It's worth noting that the MCL-MT yields BLEU score gains on zero-shot translations but at a cost of a slight decrease (approximately 0.5) in BLEU scores on English-centric transla- tions. In comparison, MCL-MT manages to im- prove zero-shot translation significantly while not compromising performance on English-centered translations. This makes MCL-MT an excellent choice for many-to-many translations, including both English-centric and non-English directions.

5. Analysis

To understand the source of these performance gains, we'll perform a number of analytical exper- iments. We'll first take a look at MCL-MT's per- formance across different scenarios, then use sen- tence representations from MCL-MT to retrieve similar sentences in different languages to validate our arguments for improving uniform linguistic representations learned by MCL-MT.

Lastly, we'll visualize the sentence representations to prove that MCL-MT does in- deed bring the representations closer together.

5.1 Ablation Study

To further explore the effectiveness of MCL-MT, we examined models of different settings and summarized the findings in Table 4:

• 2 v.s. 5 : 5 Lack of token-level loss func- tion, which is not effective in unsupervised and zero-shot cases. This makes it evident that Contrastive Loss can be used to improve zero-shot translation quality without dimin- ishing performance on other directions.

• 3 v.s. 5 : 3 did not perform well with- out sentence level. This indicates that Con- trastive Loss is essential for getting good re- sults on zero-shot translations.

• 4 v.s. 5 : 4 did not perform well without word level. This indicates that Contrastive Loss is essential for getting good results on zero-shot translations.

• 5: MCL-MT yields further improvements in BLEU scores across all three scenarios, particularly for unsupervised directions. This leads us to believe that adding a multi-level loss function to the MCL-MT model does al- low for learning a more comprehensive rep- resentation space.

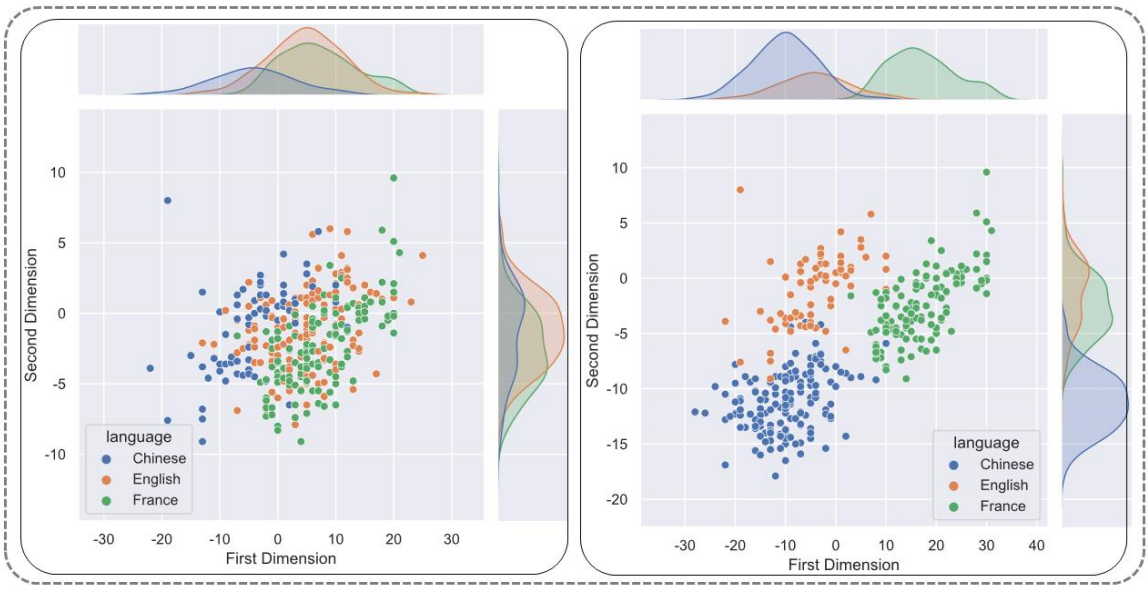

Figure 3: The figure shows bivariate kernel density estimation plots of sentence representations that have undergone a dimensionality reduction using T-SNE to 2 dimensions. The Chinese, English, and French representations are depicted by the blue, orange, and green lines, respectively. It can be observed that after implementing MCL-MT, the sentence representations are brought closer to each other.

6. Conclusion

We ultimately proved that contrastive learning at the word level can significantly improve zero-shot machine translation directions. For unsupervised tasks, our approach has achieved further advance- ment in multilingual translation. The results fur- ther demonstrated that the contrastive learning at the word level is indeed able to reduce the dispar- ity of expression between different languages. Ad- ditionally, the results also suggested that it is possible to conduct multilingual contrastive learning on word-level for machine translation.

References

[1] Maruan Al-Shedivat and Ankur P. Parikh. 2019. Con- sistency 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 Pa- pers), pages 1184–1197. Association for Computa- tional Linguistics.

[2] Ankur Bapna and Orhan Firat. 2019. Simple, scal- able adaptation for neural machine translation. InProceedings of the 2019 Conference on Empiri- cal 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.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. InProceedings of the 37th International

Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of Proceedings of Ma- chine Learning Research, pages 1597–1607. PMLR.

[4] 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 Intelli- gence, (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.

[5] 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 Associ- ation (ELRA).

[6] Chenhui Chu andRajDabre. 2019. Multilingual multi- domain adaptation approaches for neural machine translation. CoRR, abs/1906.07978.

[7] Alexis Conneau and Guillaume Lample. 2019. Cross- lingual language model pretraining. In Advances in Neural Information Processing Systems 32: An- nual Conference on Neural Information Processing

[8] USA, June 2-7, 2019, Volume 1 (Long and Short Pa- pers), pages 4171–4186. Association for Computa- tional Linguistics.

[9] Raj Dabre, Fabien Cromier[è]s, and Sadao Kurohashi. 2017. Enabling multi-source neural machine trans- lation by concatenating source sentences in multiple languages. CoRR, abs/1702.06135.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language under- standing. 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,

[11] Carlos Escolano, Marta R. Costa-juss[à], Jos[é] A. R. Fonollosa, and Mikel Artetxe. 2020. Training mul- tilingual machine translation by alternately freez- ing language-specific encoders-decoders. CoRR, abs/2006.01594.

[12] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Man- deep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vi- taliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Be- yond english-centric multilingual machine transla- tion. CoRR, abs/2010.11125.

[13] Hongchao Fang and Pengtao Xie. 2020. CERT: con- trastive self-supervised learning for language under- standing. CoRR, abs/2005.12766.

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

[15] Baijun Ji, Zhirui Zhang, Xiangyu Duan, Min Zhang, Boxing Chen, and Weihua Luo. 2020. Cross- lingual pre-training based transfer for zero-shot neural ma- chine translation. In The Thirty-Fourth AAAI Con- ference 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, Febru- ary 7-12, 2020, pages 115–122. AAAI Press