When do Contrastive Word Alignments Improve Many-to-many Neural Machine Translation?

Zhuoyuan Mao [μ] **Chenhui Chu** [μ] **Raj Dabre** [ν]
Haiyue Song [μ] **Zhen Wan** [μ] **Sadao Kurohashi** [μ]
[μ] Kyoto University, Japan [ν] NICT, Japan
{zhuoyuanmao, chu, song, zhenwan, kuro}@nlp.ist.i.kyoto-u.ac.jp
raj.dabre@nict.go.jp

Abstract

Word alignment has proven to benefit many-to-many neural machine translation (NMT). However, high-quality ground-truth bilingual dictionaries were used for pre-editing in previous methods, which are unavailable for most language pairs. Meanwhile, the contrastive objective can implicitly utilize automatically learned word alignment, which has not been explored in many-to-many NMT. This work proposes a word-level contrastive objective to leverage word alignments for many-to-many NMT. Empirical results show that this leads to 0.8 BLEU gains for several language pairs. Analyses reveal that in many-to-many NMT, the encoder's sentence retrieval performance highly correlates with the translation quality, which explains when the proposed method impacts translation. This motivates future exploration for many-to-many NMT to improve the encoder's sentence retrieval performance.

1 Introduction

Many-to-many neural machine translation (NMT) (Firat et al., 2016; Johnson et al., 2017; Aharoni et al., 2019; Sen et al., 2019; Arivazhagan et al., 2019b; Lin et al., 2020; Pan et al., 2021b) jointly trains a translation system for multiple language pairs and obtain significant gains consistently across many translation directions. Previous work (Lin et al., 2020) shows that word alignment information helps improve pre-training for many-to-many NMT. However, manually cleaned high-quality ground-truth bilingual dictionaries are used to pre-edit the source sentences, which are unavailable for most language pairs.

Recently, contrastive objectives (Clark et al., 2020; Gunel et al., 2021; Giorgi et al., 2021; Wei et al., 2021; Mao et al., 2021) have been shown to be superior at leveraging alignment knowledge in various NLP tasks by contrasting the representations of positive and negative samples in a discriminative manner. This objective, which should be able to utilize word alignment learned by any toolkit, which in turn will remove the constraints of using manually constructed dictionaries, has not been explored in the context of leveraging word alignment for many-to-many NMT.

An existing contrastive method (Pan et al., 2021b) for many-to-many NMT relies on sentence-level alignments. Given that the incorporation of word alignments has led to improvements in previous work, we believe that fine-grained contrastive objectives focusing on word alignments should help improve translation. Therefore, this paper proposes word-level contrastive learning for many-to-many NMT using the word alignment extracted by automatic aligners. We conduct experiments on three many-to-many NMT systems covering general and spoken language domains. Results show that our proposed method achieves significant gains of 0.8 BLEU in the general domain compared to previous word alignment based methods and the sentence-level contrastive method.

We then analyze how the word-level contrastive objective affects NMT training. Inspired by previous work (Artetxe and Schwenk, 2019) that train sentence retrieval models using many-to-many NMT, we speculate that our contrastive objectives affect the sentence retrieval performance and subsequently impact the translation quality. Further investigation reveals that in many-to-many NMT, the sentence retrieval precision of the multilingual encoder for a language pair strongly correlates with its translation quality (BLEU), which provides insight about when contrastive alignment improves translation. This revelation emphasizes the importance of improving the retrieval performance of the encoder for many-to-many NMT.

2 Word-level Contrastive Learning for Many-to-many NMT

Inspired by the contrastive learning framework (Chen et al., 2020) and the sentence-level con-

Findings of the Association for Computational Linguistics: NAACL 2022, pages 1766 - 1775

trastive learning objective (Pan et al., 2021b), we propose a word-level contrastive learning objective to explicitly guide the training of the multilingual encoder to obtain well-aligned cross-lingual representations. Specifically, we use word alignments, obtained using automatic word aligners, to supervise the training of the multilingual encoder by a contrastive objective alongside the NMT objective.

Alignment Extraction Two main approaches for automatically extracting aligned words from a sentence pair are: using a bilingual dictionary and using unsupervised word aligners. The former extracts fewer but precise alignments, whereas the latter extracts more but noisy alignments. We extract word-level alignments by both methods and explore how they impact NMT training. For the former approach, we use word2word (Choe et al., 2020) to construct bilingual lexicons and then extract word pairs from parallel sentences. The extracted word pairs are combined to form a phrase if words are consecutive in the source and target sentence. For the latter approach, we use FastAlign (Dyer et al., 2013) and use only 1-to-1 mappings for training.

Word-level Contrastive Learning With the extracted alignments, we propose a word-level contrastive learning objective for the multilingual encoder by the motivation that the aligned words within a sentence pair should have a similar contextual representation. We expect the supervision of the contrastive objective on the corresponding contextual word representation leads to a robust multilingual encoder. Assume that the tokenized source and target parallel sentences in the $i-th$ batch are $\mathcal{D}_i = \{src_{ij}, tgt_{ij}\}_{j=1}^{B}$, and the extracted alignments from all the sentence pairs in each batch are $\mathcal{A}_i = \{s_{ik}, t_{ik}\}_{k=1}^{N}$, where B and N denote the batch-size and the number of alignments, respectively. Note that s_{ik} and t_{ik} may contain several tokens after the word combination for word2word or subword tokenization for NMT. Then the word-level contrastive loss in a batch is:

$$\mathcal{L}_{align}^{(i)} = -\sum_{k=1}^{N}(\log \frac{\exp\left(sim(s_{ik}, t_{ik})/\mathcal{T}\right)}{\sum_{m=1}^{N}\exp\left(sim(s_{ik}, t_{im})/\mathcal{T}\right)} + \log \frac{\exp\left(sim(s_{ik}, t_{ik})/\mathcal{T}\right)}{\sum_{m=1}^{N}\exp\left(sim(s_{im}, t_{ik})/\mathcal{T}\right)}) \quad (1)$$

where $\mathcal{T}$ denotes a similarity scaling temperature. The similarity between two words is measured by:

$$sim(word_x, word_y) = cos(g(\bar{\mathbf{x}}), g(\bar{\mathbf{y}})) \quad (2)$$

where $g(\mathbf{x}) = \mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{x})$ and $\bar{\mathbf{x}}$ denotes the average of contextual hidden states of the corresponding subword positions on top of the multilingual encoder. Following (Chen et al., 2020), we use an MLP between contrastive loss and the contextual representation for NMT loss. ReLU activation is used for σ, $\mathbf{W}_1$ is $d \times d$ and $\mathbf{W}_2$ is $d \times d'$, where d is the encoder's hidden dimension and $d' < d$.

Finally, to jointly train with the NMT loss, we use the following equation to combine our proposed word-level contrastive loss for a batch:

$$\mathcal{L}^{(i)} = \frac{1}{B}(\mathcal{L}_{NMT}^{(i)} + w\frac{N_T}{2N}\mathcal{L}_{align}^{(i)}) \quad (3)$$

where N_T is the number of the tokens within a batch, $\frac{N_T}{2N}$ is a multiplier that scales the contrastive loss to be consistent with NMT loss, and w is a weight to balance the joint training.

La. pair	Source	Size	N (w2w)	N (FA)
en-et	WMT18	1.9M	5,762,977	38,454,477
en-it	IWSLT17	231k	603,032	3,000,011
en-ja	IWSLT17	223k	684,583	2,797,882
en-kk	WMT19	124k	124,511	279,429
en-my	ALT	18k	75,383	377,392
en-nl	IWSLT17	237k	564,697	2,836,873
en-ro	WMT16	612k	3,271,848	13,092,240
en-tr	WMT17	207k	770,873	2,885,102
en-vi	IWSLT15	133k	354,167	2,120,755

Table 1: **Data Source and number of the extracted word pairs.** La. pair, N (w2w) and N (FA) denote the language pair, the number of the word pairs extracted by word2word and FastAlign, respectively. Refer to Appendix B for details of the dataset splits.

3 Experimental Settings

Datasets and Preprocessing We selected ten languages, including English (en), Estonian (et), Italian (it), Japanese (ja), Kazakh (kk), Burmese (my), Dutch (nl), Romanian (ro), Turkish (tr), Vietnamese (vi) from different language families to train the NMT systems. We used the parallel datasets from different domains for the selected nine language pairs, including IWSLT, WMT, and ALT. We followed mBART (Liu et al., 2020) for tokenization. Details are given in Appendix A. For each parallel dataset, we implemented two approaches as stated in Section 2 to extract word pairs for the contrastive training objective. Data source and the number of the extracted word pairs are shown in Table 1. To ensure high alignment

Methods	en-tr		en-ro		en-et		en-kk		en-my	
	→	←	→	←	→	←	→	←	→	←
MLSC	9.3	12.6	25.0	26.2	10.8	15.1	0.5	5.3	15.1	15.6
+align	9.0	12.4	24.6	26.5	10.7	14.6	0.4	5.4	15.0	15.3
+w2w (ours)	9.4	12.6	24.8	26.8	10.8	15.1	0.5	5.8	15.2	15.9
+FA (ours)	9.1	12.2	24.8	26.7	10.7	14.8	0.3	5.6	15.0	15.6
mBART FT	17.7	22.2	33.8	37.1	14.5	24.3	1.8	14.1	17.8	23.1
+align	17.5	21.9	33.8	36.7	15.2	24.3	1.8	14.0	16.9	22.1
+w2w (ours)	17.6	22.2	34.2	37.5	15.0	25.0	1.2	14.1	18.3	23.8
+FA (ours)	17.5	22.2	34.3	37.5	14.9	25.1	1.3	14.4	17.9	23.6

Table 2: **BLEU scores of 626_en-tr-ro-et-my-kk system.** Significantly better scores (Koehn, 2004) are in cyan, and marginal improvements are in lightcyan.

Methods	222_en-ja	626_I	626_II
MLSC	13.90	23.76	13.55
+align	13.90	23.67	13.39
+w2w (ours)	13.85	23.44	**13.69**
+FA (ours)	13.30	23.68	13.48
mBART FT	18.90	29.11	20.64
+align	18.55	28.87	20.42
+w2w (ours)	18.80	29.08	**20.89**
+FA (ours)	18.65	29.01	**20.87**

Table 3: **Overall average BLEU of all the systems.** 626_I and 626_II denote 626_en-it-ja-nl-tr-vi and 626_en-tr-ro-et-my-kk, respectively. Results better than MLSC or mBART FT are marked **bold**. Refer to Appendix D for the detailed scores of all the systems.

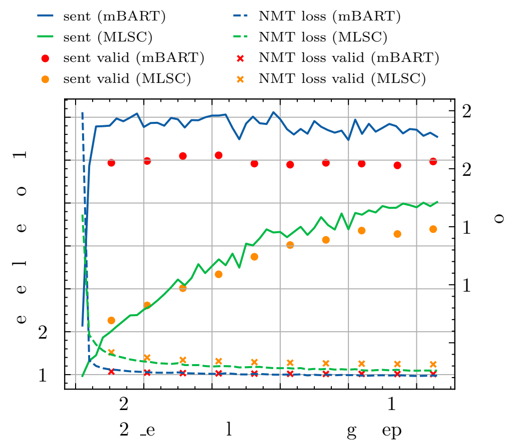

Figure 1: **NMT loss, sentence retrieval P@1 of the encoder in MLSC and mBART FT.** The average of the contextual embeddings on top of the encoder is used as the sentence embedding. We report the average in-batch retrieval precision of both directions of each language pair.

quality, we used large-scale out-of-domain (see Appendix B) parallel corpora with FastAlign.

Many-to-many NMT systems We established three many-to-many NMT systems as follows:

222_en-ja: Bidirectional en-ja NMT model using en-ja parallel corpus.

626_en-it-ja-nl-tr-vi: 6-to-6 multilingual NMT model using spoken language domain corpora for en-it, en-ja, en-nl, en-tr and en-vi.

626_en-tr-ro-et-my-kk: 6-to-6 multilingual NMT model using general domain corpora for en-tr, en-ro, en-et, en-my and en-kk.

Baselines and Ours For each language group setting above, we conducted NMT experiments on both the multilingual training from scratch (**MLSC**) (Johnson et al., 2017; Aharoni et al., 2019) and the mBART multilingual fine-tuning (**mBART FT**) (Tang et al., 2020) as baselines. We applied our proposed word-level contrastive learning in both MLSC and mBART FT, and compared with another strong baseline, word alignment based joint NMT training (**+align**) (Garg et al., 2019). For applying our method, we investigated the performance of joint training with word pairs extracted by both word2word (**+w2w**) and FastAlign (**+FA**). We omitted Lin et al. (2020) as a baseline because their method can not be applied to mBART fine-tuning, and they used high-quality ground-truth dictionaries, which are unavailable for most languages pairs.

Implementation We used `mBART-large` (mBART-25) for mBART FT and `transformer-base` (Vaswani et al., 2017) for MLSC. See Appendix C for details.

4 Results and Analyses

BLEU Results We report case-sensitive tokenized BLEU (Papineni et al., 2002) results in Table 3 and 2. In Table 3, we observe that with our pro-

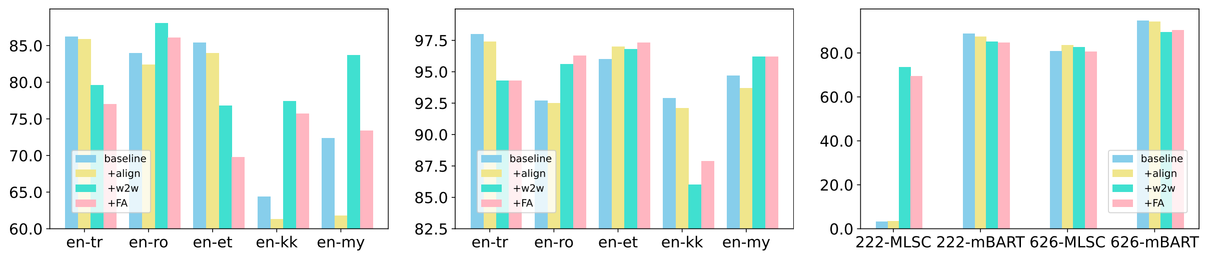

Figure 2: **Sentence retrieval P@1 on the validation set for each language pair.** *Left* and *middle* are the results on 626_en-tr-ro-et-my-kk MLSC and mBART FT, respectively. "626" in *right* subfigure denote 626_en-it-ja-nl-tr-vi. Refer to Appendix E for setup and results in details.

posed training objectives, BLEU scores are comparable in 222_en-ja and 626_en-it-ja-nl-tr-vi while they are slightly improved in 626_en-tr-ro-et-my-kk. However, "+align" performs comparable or even worse compared with the baseline. Referring to Table 2 for specific BLEUs on each language pair, we find that with our methods, translation performances are significantly improved for mBART FT while nontrivial improvements can merely be observed on en-ro and en-kk direction for MLSC. This indicates that NMT fine-tuning on monolingual pre-trained models (mBART) may benefit more from our proposed methods. Note that the BLEU improvements for MLSC are not significant, and we explain why this happens in the "Word Retrieval P@1 is improved" part.

Latent Encoder Alignment Property We now inspect which aspect of alignment-based methods impacts the translation performance. Previous work (Artetxe and Schwenk, 2019) showed that the encoder of a strong multilingual NMT system is an ideal model for the bilingual sentence retrieval task. In addition, Arivazhagan et al. (2019a) introduced the correlation between the encoder-side sentence representation[1] and the translation quality. Inspired by these, we speculate that alignment-based objectives affect sentence retrieval performance, which further impacts the translation quality. We train MLSC and mBART FT and report the sentence retrieval precision and NMT loss during the training. Results are reported in Figure 1. We observe that the validation retrieval precision show similar trends as the NMT loss. This indicates that during many-to-many NMT training from scratch, encoder-side sentence-level retrieval precision is optimized along with the NMT loss.

Sentence Retrieval P@1 Correlates with BLEU According to the investigation of the encoder alignment property above, we verify the relationship between BLEU score and sentence retrieval precision on the validation set for each language pair. Results are shown in Figure 2. Cross-referencing the BLEU score in Table 2, we found that BLEU scores are improved when the encoder achieves gains on the sentence retrieval precision.[2] For example, we see increases of the retrieval P@1 on en-ro, en-et, and en-my on mBART FT (the middle of Figure 2) while BLEU scores are significantly improved on these three language pairs (Table 2). We further calculate the Pearson correlation coefficient between the BLEU changes and sentence retrieval P@1 changes for mBART+align, mBART+w2w, and mBART+FA in the 626_en-tr-ro-et-my-kk setting. Results are 0.79, 0.93, 0.90, respectively, demonstrating a strong correlation between translation quality and sentence retrieval precision.

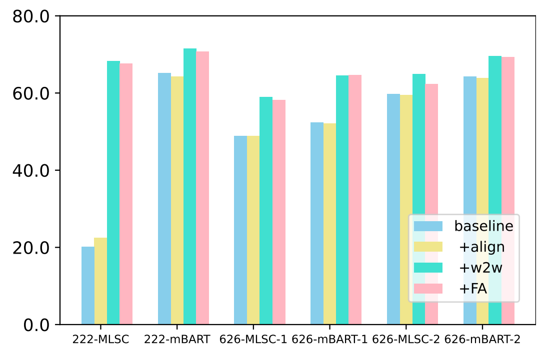

Figure 3: **Average Word retrieval P@1 on the validation set for each language pair.** "626-*-1" and "626-*-2" indicate 626_en-it-ja-nl-tr-vi and 626_en-tr-ro-et-my-kk, respectively. Refer to Appendix F for setup and results in details.

Word Retrieval P@1 is Improved We probe the trained contextualized word representations on top of the encoder. As shown in Figure 3, we observe that the word retrieval precision is improved in all

[1] Usually a pooled encoder output.

[2] 222_en-ja MLSC setting can hardly learn a well-aligned encoder while our methods improve the encoder sentence-level alignment quality without sacrificing BLEU scores.

the settings. This demonstrates that the encoder parameters of the NMT system trained with our proposed objective are of a rather different distribution. By just changing the random seed, we can expect similar BLEU results, but we cannot obtain a better aligned encoder. However, the improvement of the word retrieval precision does not directly contribute to the translation quality, which we explain next.

Word-level Contrastive Objective and Sentence Retrieval P@1 With the word-level contrastive objective, we observed significant BLEU score improvements on language pairs such as en-ro, en-et and en-my for mBART FT as presented in Table 2. However, noisy word pairs (Pan et al., 2021a) extracted via word alignment toolkits leads to poor supervision signals for improving sentence retrieval P@1, which in turn prevents some language pairs such as en-kk from exhibiting BLEU improvements. We found that for en-kk, the numbers of extracted word pairs per sentence by word2word and FastAlign are 1.0 and 2.2, respectively. In contrast, these numbers are 4.2 and 20.7 for improved language pairs, calculated from Table 1. Although better extracted word alignments for the word-level contrastive objective leads to BLEU improvements, its contribution towards improvements varies for MLSC and mBART FT, as shown in Table 2. We expect these findings to provide new perspectives for improving many-to-many NMT.

Sentence-level Contrastive Objective We conducted the experiments for the sentence-level contrastive objective (Pan et al., 2021b) on all two six-to-six settings and compared it against our proposed approach. The average BLEUs of our methods significantly outperform those of sentence-level contrastive objectives (see Table 8 and 9), clearly showing the sentence-level objective's limitation. Moreover, we checked the sentence retrieval P@1 for Pan et al. (2021b) (Table 10 and 11) and found that it correlates with BLEU changes, indicating that sentence-level contrastive objective is suboptimal for language pairs with decreased retrieval precision.[3]

5 Conclusion

We proposed a word-level contrastive learning objective for many-to-many NMT. Experimental results showed that our proposed method leads to significantly better translation for several language pairs, which is then explained by analyses showing the relationship between BLEU scores and sentence retrieval performance of the NMT encoder. Future work can focus on: (1) further improving the encoder's retrieval performance in many-to-many NMT; (2) contrastive objective's feasibility in a massively multilingual scenario.

[3]Note that the sentence-level contrastive objective incorporates sentences in multiple languages for contrastive loss. It does not necessarily improve the pair-wise retrieval precision.

Ethical Considerations

All the corpora we used in this paper are publicly available resources without the issue of the copyright. The technique this paper proposed is for NMT models, so it can not circumvent the issues that NMT models have. Since our automatically dictionaries are extracted from potentially biased data, the translations may also contain biases. However, we expect that these issues may be resolved by using unbiased data or the addition of debiasing objectives.

Acknowledgements

This work was supported by Grant-in-Aid for Young Scientists #19K20343, JSPS and Information/AI/Data Science Doctoral Fellowship of Kyoto University.

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.

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

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. 2019b. 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. *Transactions of the Association for Computational Linguistics*, 7:597–610.

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.

Yo Joong Choe, Kyubyong Park, and Dongwoo Kim. 2020. word2word: A collection of bilingual lexicons for 3,564 language pairs. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3036–3045, Marseille, France. European Language Resources Association.

Christos Christodoulopoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Lang. Resour. Evaluation*, 49(2):375–395.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Chenchen Ding, Hnin Thu Zar Aye, Win Pa Pa, Khin Thandar Nwet, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita. 2020. Towards burmese (myanmar) morphological analysis: Syllable-based tokenization and part-of-speech tagging. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 19(1):5:1–5:34.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

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.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. 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)*, pages 4453–4462, Hong Kong, China. Association for Computational Linguistics.

John M. Giorgi, Osvald Nitski, Bo Wang, and Gary D. Bader. 2021. Declutr: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 879–895. Association for Computational Linguistics.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. 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.

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. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Zhuoyuan Mao, Prakhar Gupta, Chenhui Chu, Martin Jaggi, and Sadao Kurohashi. 2021. Lightweight cross-lingual sentence representation learning. 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 2902–2913, Online. Association for Computational Linguistics.

Hajime Morita, Daisuke Kawahara, and Sadao Kurohashi. 2015. Morphological analysis for unsegmented languages using recurrent neural network language model. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language*

Processing, pages 2292–2297, Lisbon, Portugal. Association for Computational Linguistics.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2204–2208, Portorož, Slovenia. European Language Resources Association (ELRA).

Lin Pan, Chung-Wei Hang, Haode Qi, Abhishek Shah, Saloni Potdar, and Mo Yu. 2021a. Multilingual BERT post-pretraining alignment. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 210–219, Online. Association for Computational Linguistics.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021b. 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.

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multilingual unsupervised NMT using shared encoder and language-specific decoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3083–3089, Florence, Italy. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401.

Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. 2018. Juman++: A morphological analysis toolkit for scriptio continua. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 54–59, Brussels, Belgium. 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.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2021. On learning universal representations across languages. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

La. pair	Train	Valid	Test	OD Size
en-et	WMT18	WMT18	WMT18	10.7M
en-it	IWSLT17	IWSLT15	IWSLT16	13.6M
en-ja	IWSLT17	IWSLT15	IWSLT16	10.7M
en-kk	WMT19	WMT19	WMT19	851k
en-my	ALT	ALT	ALT	446k
en-nl	IWSLT17	IWSLT15	IWSLT16	12.7M
en-ro	WMT16	WWT16	WMT16	11.0M
en-tr	WMT17	WWT16	WMT16	11.1M
en-vi	IWSLT15	IWSLT13	IWSLT14	11.9M

Table 4: **Dataset statistics for each language pair.** "La. pair" means language pair and "OD Size" denotes the number of the out-of-domain sentence pairs used for training FastAlign.

Methods	en-ja	ja-en
MLSC	15.9	11.9
+align	16.3	11.5
+w2w (ours)	16.0	11.7
+FA (ours)	15.6	11.0
mBART FT	19.8	18.0
+align	19.6	17.5
+w2w (ours)	19.4	18.2
+FA (ours)	19.5	17.8

Table 5: **BLEU scores of 222_en-ja system.** Significantly better scores are in cyan, and marginal improvements are in lightcyan. The significance test is done with Koehn (2004).

Methods	en-ja
MLSC	3.3
+align	**3.5**
+w2w (ours)	**73.5**
+FA (ours)	**69.6**
mBART FT	88.9
+align	87.4
+w2w (ours)	85.2
+FA (ours)	84.8

Table 6: **Sentence retrieval P@1 on the validation set for 222_en-ja.**

Methods	en-ja
MLSC	20.1
+align	**22.5**
+w2w (ours)	**68.3**
+FA (ours)	**67.6**
mBART FT	65.2
+align	64.3
+w2w (ours)	**71.5**
+FA (ours)	**70.7**

Table 7: **Word retrieval P@1 on the validation set for 222_en-ja.**

A Tokenization Settings

For Japanese, we use Jumanpp (Morita et al., 2015; Tolmachev et al., 2018) for segmentation, and we follow the same settings as in mBART (Liu et al., 2020) for other languages: `myseg.py` (Ding et al., 2020) is used for Burmese, Moses tokenization and special normalization is used for Romanian following (Sennrich et al., 2016),[4] and Moses tokenization for other languages.[5] Following mBART, we apply SentencePiece (Kudo and Richardson, 2018) to further segment sentences into subwords.[6]

B Datasets and Alignment Extraction

The datasets used for NMT training, validation and test are shown in Table 4. For the word alignment extraction using FastAlign, we also use out-of-domain parallel corpora to train the FastAlign jointly, aiming to obtain word alignments with less noise. The out-of-domain corpora for all the language pairs contain Tatoeba, Europarl, GlobalVoices, NewsCommentary, OpenSubtitles, TED, WikiMatrix, QED, GNOME, bible-uedin, and ASPEC (Nakazawa et al., 2016). We collect them from the OPUS project (Christodoulopoulos and Steedman, 2015) and WAT.[7] The number of the out-of-domain parallel sentences for each language pair is shown in Table 4.

C Implementation Details

Following Tang et al. (2020), we set the oversampling temperature of 1.5 for all the settings. For MLSC, we set the dropout of 0.3 to avoid overfitting on small-scale training data. We used the batch size of 1,024 tokens for all the settings. For our word-level contrastive learning, we set the weight of 0.1, the temperature of 0.2, d' of 128, and a

[4] https://github.com/rsennrich/wmt16-scripts

[5] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

[6] https://github.com/google/sentencepiece

[7] https://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2021/index.html

Methods	en-ja		en-vi		en-it		en-nl		en-tr		Avg.
	→	←	→	←	→	←	→	←	→	←	
MLSC	15.4	11.8	29.6	28.6	27.5	32.7	29.1	36.4	11.6	14.9	23.76
+align	15.1	11.4	29.4	28.3	27.7	33.0	28.9	36.0	11.8	15.1	23.67
+w2w (ours)	15.3	11.6	29.7	28.2	27.6	32.4	28.6	35.8	10.8	14.4	23.44
+FA (ours)	15.5	11.6	29.6	28.0	27.8	33.2	29.1	35.9	11.2	14.9	23.68
+sent	15.1	11.6	29.6	28.3	27.3	32.7	28.1	36.6	11.3	14.7	23.53
mBART FT	17.8	17.0	34.1	35.7	32.5	38.0	32.6	41.6	18.7	23.1	29.11
+align	17.6	16.7	33.7	35.6	32.0	37.7	32.5	41.3	18.7	22.9	28.87
+w2w (ours)	17.6	17.2	34.2	35.7	32.5	38.2	32.1	41.7	18.7	22.9	29.08
+FA (ours)	17.5	17.7	34.0	35.2	32.4	37.9	32.3	41.4	18.6	23.1	29.01
+sent	17.8	16.5	33.7	35.6	32.2	38.1	32.5	41.2	18.1	22.9	28.86

Table 8: **BLEU scores of 626_en-it-ja-nl-tr-vi system.** Significantly better scores are in cyan, and marginal improvements are in lightcyan. The significance test is done with Koehn (2004).

Methods	en-tr		en-ro		en-et		en-kk		en-my		Avg.
	→	←	→	←	→	←	→	←	→	←	
MLSC	9.3	12.6	25.0	26.2	10.8	15.1	0.5	5.3	15.1	15.6	13.55
+align	9.0	12.4	24.6	26.5	10.7	14.6	0.4	5.4	15.0	15.3	13.39
+w2w (ours)	9.4	12.6	24.8	26.8	10.8	15.1	0.5	5.8	15.2	15.9	13.69
+FA (ours)	9.1	12.2	24.8	26.7	10.7	14.8	0.3	5.6	15.0	15.6	13.48
+sent	8.7	12.1	24.5	26.0	10.4	14.5	0.4	5.3	13.8	14.6	13.03
mBART FT	17.7	22.2	33.8	37.1	14.5	24.3	1.8	14.1	17.8	23.1	20.64
+align	17.5	21.9	33.8	36.7	15.2	24.3	1.8	14.0	16.9	22.1	20.42
+w2w (ours)	17.6	22.2	34.2	37.5	15.0	25.0	1.2	14.1	18.3	23.8	20.89
+FA (ours)	17.5	22.2	34.3	37.5	14.9	25.1	1.3	14.4	17.9	23.6	20.87
+sent	17.2	22.1	34.2	37.0	14.2	24.1	1.6	14.0	17.7	23.4	20.55

Table 9: **BLEU scores of 626_en-tr-ro-et-my-kk system.** Significantly better scores are in cyan, and marginal improvements are in lightcyan. The significance test is done with Koehn (2004).

Methods	en-ja	en-vi	en-it	en-nl	en-tr	Avg.
MLSC	52.7	84.6	91.0	85.7	89.7	80.9
+align	**53.5**	82.8	**91.2**	**86.4**	88.9	80.6
+w2w (ours)	**73.4**	**85.7**	**91.4**	84.7	83.1	**83.7**
+FA (ours)	**71.3**	**84.9**	**91.3**	83.8	82.0	**82.7**
+sent	87.2	84.7	91.1	87.7	86.6	87.5
mBART FT	87.1	96.2	97.3	94.6	98.5	94.7
+align	85.1	95.8	97.3	94.2	98.5	94.2
+w2w (ours)	81.6	91.4	94.7	90.8	89.6	89.6
+FA (ours)	82.6	92.3	95.0	91.7	90.4	90.4
+sent	76.2	88.3	93.6	88.7	89.8	87.3

Table 10: **Sentence retrieval P@1 on the validation set for 626_en-it-ja-nl-tr-vi.**

Methods	en-tr	en-ro	en-et	en-kk	en-my	Avg.
MLSC	86.2	84.0	85.4	64.4	72.4	78.5
+align	85.9	82.4	84.0	61.3	61.8	75.1
+w2w (ours)	79.6	**88.1**	76.8	**77.4**	**83.7**	**81.1**
+FA (ours)	77.0	**86.1**	69.8	**75.7**	**73.4**	76.4
+sent	76.3	77.6	55.2	63.8	71.4	68.9
mBART FT	98.0	92.7	96.0	92.9	94.7	94.9
+align	97.4	92.5	**97.0**	92.1	93.7	94.5
+w2w (ours)	94.3	**95.6**	**96.8**	86.0	**96.2**	93.8
+FA (ours)	94.3	**96.3**	**97.3**	87.9	**96.2**	94.4
+sent	94.6	**97.3**	95.4	**93.1**	**95.7**	**95.2**

Table 11: **Sentence retrieval P@1 on the validation set for 626_en-tr-ro-et-my-kk.**

smaller dropout of 0.2 because our proposed objective serves as a regularization part. We followed the hyperparameter setting of Garg et al. (2019) for word alignment-based joint NMT training. We used 8 NVIDIA A100 for mBART FT and 8 TITAN Xp for MLSC model training. The model is validated every 1000 steps for 222_en-ja and 2000 steps for both two 626 settings. We do the early

Methods	en-ja	en-vi	en-it	en-nl	en-tr	Avg.
MLSC	61.8	54.6	42.8	42.1	42.7	48.8
+align	**61.9**	54.1	**43.7**	42.0	42.3	48.8
+w2w (ours)	**64.0**	**64.7**	**55.8**	**57.7**	**52.8**	**59.0**
+FA (ours)	58.2	**65.2**	**59.2**	**60.1**	**48.1**	**58.2**
mBART FT	64.5	57.2	47.4	45.9	47.2	52.4
+align	64.0	56.8	47.3	45.7	46.8	52.1
+w2w (ours)	**71.3**	**70.1**	**60.6**	**62.9**	**57.8**	**64.5**
+FA (ours)	**68.6**	**69.4**	**63.2**	**64.7**	**57.4**	**64.7**

Table 12: **Word retrieval P@1 on the validation set for 626_en-it-ja-nl-tr-vi.**

Methods	en-tr	en-ro	en-et	en-kk	en-my	Avg.
MLSC	41.9	63.2	64.4	63.4	65.8	59.7
+align	40.9	63.2	63.9	63.4	**66.2**	59.5
+w2w (ours)	**50.1**	**66.5**	**67.6**	**68.8**	**71.3**	**64.9**
+FA (ours)	**47.2**	**66.7**	**65.7**	**65.4**	**66.3**	**62.3**
mBART FT	46.8	66.1	68.0	68.7	71.7	64.3
+align	46.4	65.9	67.8	68.5	71.1	63.9
+w2w (ours)	**55.6**	**70.3**	**72.8**	**74.7**	**74.4**	**69.6**
+FA (ours)	**55.3**	**70.1**	**73.0**	**74.0**	**74.0**	**69.3**

Table 13: **Word retrieval P@1 on the validation set for 626_en-tr-ro-et-my-kk.**

stopping if no improvement of the validation loss is observed for 8 checkpoints. The model with the best validation loss was used for evaluation.

D BLEU Scores

We report all the BLEU results of 222_en-ja, 626_en-it-ja-nl-tr-vi, and 626_en-tr-ro-et-my-kk in Table 5, 8 and 9, respectively.

E Sentence Retrieval Precision

We report the sentence retrieval precision for all the systems in Tables 6, 10 and 11. The sentence retrieval previsions are evaluated by using the validation dataset of each language pair. The mean pooled encoder output is used as the sentence embedding. We use cosine similarity to conduct the retrieval task, and report the average retrieval precision of both directions of each language pair.

F Word Retrieval Precision

We report the word retrieval precision for all the systems in Tables 7, 12, and 13. The word retrieval precision are computed by using the validation dataset and the word2word alignments on it. The mean pooled encoder output on corresponding positions is used as the contextualized word embedding. We use cosine similarity to implement the retrieval for word pairs in a batch, and present the average in-batch retrieval precision of both directions of each language pair. Batch size is set as 512 tokens.