Appendix: Refining Low-Resource Unsupervised Translation by Language Disentanglement of Multilingual Model

Xuan-Phi Nguyen[1,3], **Shafiq Joty**[1,2], **Wu Kui**[3], **Ai Ti Aw**[3]
[1]Nanyang Technological University
[2]Salesforce Research
[3]Institute for Infocomm Research (I^2R), A*STAR
Singapore
{nguyenxu002@e.ntu,srjoty@ntu}.edu.sg
{wuk,aaiti}@i2r.a-star.edu.sg

A Appendix

In the Appendix, we provide further details about the statistics of the datasets used for training and evaluation, as well as the experimental setups. We also conduct analyses experiments into the reason why further pseudo-parallel data mining fails to improve low-resource unsupervised translation.

A.1 Statistics of Datasets and Experimental Setups

Data statistics. Table 1 provides details about the amount of monolingual data for each corpus of different languages that are tested in this paper. Apart from English, Hindi, Finnish, the amounts available to be trained for other low-resource languages are particularly low, making improving rather more difficult. Table 2, meanwhile, provide the origins of the test sets used to evaluate the performances of the models on the low-resource tasks. The test set sources are the same as used in CRISS paper [9].

Table 1: Statistics on the sizes of monolingual corpora that are used in the paper.

	No. Sentences
English (En)	100M
Nepali (Ne)	9.8M
Sinhala (Si)	8.8M
Hindi (Hi)	80M
Gujarati (Gu)	5.6M
Finnish (Fi)	61M
Latvian (Lv)	35M
Estonian (Et)	26M
Kazakh (Kk)	33M

Experimental details. For all low-resource experiments, we use the same architectural setups as CRISS [9]. That is, we use a Transformer with 12 layers, dimension of 1024 with 16 attention heads. We finetune the models with cross-entropy loss with 0.2 label smoothing, 0.3 dropout, 2500 warm-up steps. We use a learning rate of 3e-5. When decoding, we use beam size of 5 and length penalty of 0.1 for Indic languages and 0.5 for other languages.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Table 2: Test sets used for each language pair.

	Test set
En-Ne	FLoRes [5]
En-Si	FLoRes [5]
En-Hi	ITTB [6]
En-Gu	WMT19 [2]
En-Fi	WMT17 [3]
En-Lv	WMT17 [3]
En-Et	WMT18 [4]
En-Kk	WMT19 [2]

Table 3: Comparison of BLEU scores for different methods on fully unsupervised translation tasks of various low-resource languages from the Indic, Uralic and Turkic language families. SacreBLEU [8] numbers are reported in subscript.

Method	**Indic**								**Uralic**						**Turkic**	
	En-Ne	**Ne-En**	**En-Si**	**Si-En**	**En-Hi**	**Hi-En**	**En-Gu**	**Gu-En**	**En-Fi**	**Fi-En**	**En-Lv**	**Lv-En**	**En-Et**	**Et-En**	**En-Kk**	**Kk-En**
mBART	4.4	10.0	3.9	8.2	-	-	-	-	-	-	-	-	-	-	-	-
LAgSwAV	5.3$_{5.4}$	12.8$_{12.5}$	5.4$_{5.3}$	9.4$_{5.1}$	-	-	-	-	-	-	-	-	-	-	-	-
CRISS	5.5$_{5.6}$	14.5$_{14.4}$	6.0$_{6.0}$	14.5$_{14.3}$	19.4$_{19.5}$	23.6$_{23.1}$	14.2$_{14.2}$	23.7$_{23.4}$	20.2$_{20.1}$	26.7$_{26.2}$	14.4$_{14.3}$	19.2$_{18.6}$	16.8$_{16.7}$	25.0$_{24.6}$	6.7$_{6.7}$	14.5$_{14.5}$
Ours	**9.0**$_{9.1}$	**18.2**$_{18.1}$	**9.5**$_{9.4}$	**15.3**$_{15.1}$	**20.8**$_{20.9}$	**23.8**$_{23.4}$	**17.5**$_{17.5}$	**29.5**$_{29.1}$	**22.9**$_{22.8}$	**28.2**$_{27.8}$	**18.5**$_{18.2}$	**19.3**$_{18.9}$	**21.0**$_{20.9}$	**25.7**$_{25.4}$	**10.0**$_{10.0}$	**20.0**$_{19.7}$
$+\Delta$	3.5	3.7	3.5	0.8	1.4	0.2	3.3	5.8	2.7	1.5	4.1	0.1	4.2	0.7	3.3	5.5

A.2 Low-resource Unsupervised Translation Results in SacreBLEU

Because the compared baselines [9, 7] reported multi-bleu.perl scores, **??** in the main paper is also reported with multi-bleu.perl scores for consistent purpose, which are carefully calibrated to match the exact evaluation pipeline (test sets, tokenization, scripts, pre-trained models, etc) used in the previous work. In this section, we present the corresponding SacreBLEU [8] scores to provide better insights and references. As shown in Table 3, SacreBLEU scores deviate from multi-bleu.perl scores by a negligible amount or up to 0.5 BLEU, depending the tested language pair.

A.3 Extra Ablation Study

In addition to the ablation study in the main paper, we also conduct further analysis on the impact of the σ hyper-parameter, which dictates how many FFN layers in the original Transformer are replaced with our language-specific sharded FFN layers. Specifically, for every layer j of an H-layer Transformer decoder that j is divisible by σ and $1 \leq j \leq H$, we replace it with our language specific FFN layer. As a result, as σ increases, the number of replaced FFNs decreases.

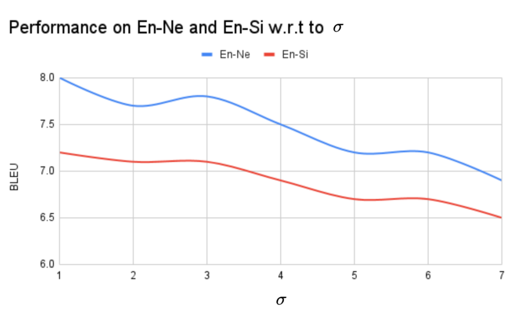

Figure 1: BLEU performance on unsupervised machine translation for En-Ne and En-Si with respect to σ. In our main experiments, we chose $\sigma = 3$ because it offers relative optimal performance while requires the least extra parameters for stage 1 to 3.

A.4 Pseudo-parallel Data Mining Analysis

Pseudo-parallel data mining has become a crucial element of unsupervised machine translation [9, 7]. CRISS [9], which we used extensively in our paper, is also itself a pseudo-parallel data mining method. This strategy involves first building a shared multilingual encoder that is able to produce language-agnostic representations of sentences across different languages. This means the model is able to map sentences of similar semantic meanings into the same latent space, regardless of which language they belong to. These latent representations can be used by LASER [1] to find and map monolingual data from two corpora together to create a synthetic parallel dataset.

In this section, given its success, we investigate whether pseudo-parallel mining can further improve the performance in low-resource unsupervised machine translation. Specifically, we use pre-trained CRISS to mine pseudo-parallel data from the monolingual corpora of English against Indic low-resource languages Nepali and Sinhala. Then, we use such data to finetune CRISS for 5000 steps (CRISS + mined data) for each low-resource pair, which is equivalent to how CRISS was originally trained, except with mined data from 180 other directions. Since the size of these mined datasets are small, we also finetune CRISS with them along with iterative back-translation on the target language pairs (CRISS + mined data + BT).

The results are reported in Table 4. Specifically, with back-translation, the mined datasets cause detrimental damage to the model's performance, mainly due to their unbearably small sizes (less than 5000 samples). Further inspections into these mined datasets show that they contains lots of mismatches or incorrect alignments, which adds more noise to the model. When combined with iterative back-translation (CRISS + mined data + BT), the model manages to beat the baseline, although the results indicates that the gains are thanks to back-translation instead of the mined data.

Table 4: Comparison with popular unsupervised MT techniques, such as CBD or CRISS pseudo-parallel data mining.

	En-Ne	Ne-En	En-Si	Si-En
Data information				
Corpus size	En: 100M, Ne: 9.8M		En: 100M, Si: 8.8M	
Mined data size	∼4800		∼3600	
Performance				
CRISS	5.5	14.5	6.0	14.5
CRISS + mined data	4.6	10.5	4.8	9.8
CRISS + mined data + BT	6.6	16.1	6.7	13.1
Ours	9.0	18.2	9.5	15.3

Upon seeing the above results, it is natural to doubt that CRISS is simply unable to mine more and better data, and a better language-agnostic encoder can mine more data. However, we empirically found that this may not be the case. To be more precise, we use our outperforming model after the proposed refinement to back-translate all sentences from the English corpus into the target low-resource language, such as Nepali (Ne). Then, for each of the back-translated Ne sentence, we use LASER to search for top 20 nearest neighbors in the real Nepali corpus based on their CRISS embeddings. We then choose the best match by the shortest Levenshtein distance between the back-translated sentence and the real ones. We perform a similar procedure for the opposite directions. After this process, we obtain a large pseudo-parallel dataset by pairing the best real Nepali sentence with the English input, which we filter out samples whose Levenshtein distances are more than 20% the average sentence length. This means we only accept sentence pairs with very low edit distance.

The sizes of these filtered mined datasets are shown in Table 5. As it can be seen, although the total unfiltered mined datasets are huge at more than 100M pairs, the filtered ones only contains less than 5000 samples for both En-Ne and En-Si. This indicates that because the monolingual corpora between English and low-resource languages are too distant and out-of-domain, there are simply not enough high quality pseudo-parallel data to be mined at all.

Table 5: Back-translated mined pseudo-parallel dataset sizes for English-Nepali and English-Sinhala.

	En-Ne	**Ne-En**	**En-Si**	**Si-En**
Data information				
Corpus size	En: 100M, Ne: 9.8M		En: 100M, Si: 8.8M	
Unfiltered mined size	~109.8M		~108.8M	
Filtered mined size	~3000		~6000	

References

[1] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, March 2019. doi: 10.1162/tacl_a_00288. URL `https://aclanthology.org/Q19-1038`.

[2] Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5301. URL `https://aclanthology.org/W19-5301`.

[3] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4717. URL `https://aclanthology.org/W17-4717`.

[4] Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6401. URL `https://aclanthology.org/W18-6401`.

[5] Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. 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 6098–6111, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1632. URL `https://www.aclweb.org/anthology/D19-1632`.

[6] Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL `https://aclanthology.org/L18-1548`.

[7] Xuan-Phi Nguyen, Hongyu Gong, Yun Tang, Changhan Wang, Philipp Koehn, and Shafiq Joty. Contrastive clustering to mine pseudo parallel data for unsupervised translation. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=pN1JOdrSY9`.

[8] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL `https://www.aclweb.org/anthology/W18-6319`.

[9] Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. Cross-lingual retrieval for iterative self-supervised training. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2207–2219. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/1763ea5a7e72dd7ee64073c2dda7a7a8-Paper.pdf`.