Disentangling Pretrained Representation to Leverage Low-Resource Languages in Multilingual Machine Translation

Frederikus Hudi[1,2], Zhi Qu[1], Hidetaka Kamigaito[1], Taro Watanabe[1]
[1]Nara Institute of Science and Technology, Nara, Japan
[2]WAP Tokushima Laboratory of AI & NLP, Tokushima, Japan
{frederikus.hudi.fe7, qu.zhi.pv5, kamigaito.h, taro}@is.naist.jp

Abstract

Multilingual neural machine translation aims to encapsulate multiple languages into a single model. However, it requires an enormous dataset, leaving the low-resource language (LRL) underdeveloped. As LRLs may benefit from shared knowledge of multilingual representation, we aspire to find effective ways to integrate unseen languages in a pre-trained model. Nevertheless, the intricacy of shared representation among languages hinders its full utilisation. To resolve this problem, we employed target language prediction and a central language-aware layer to improve representation in integrating LRLs. Focusing on improving LRLs in the linguistically diverse country of Indonesia, we evaluated five languages using a parallel corpus of 1,000 instances each, with experimental results measured by BLEU showing zero-shot improvement of 7.4 from the baseline score of 7.1 to a score of 15.5 at best. Further analysis showed that the gains in performance are attributed more to the disentanglement of multilingual representation in the encoder with the shift of the target language-specific representation in the decoder.

Keywords: Multilinguality, Machine Translation, Less-Resourced/Endangered Languages, Neural language representation models

1. Introduction

Multilingual Neural Machine Translation (MNMT) system offers efficiency from its capability to handle multiple language pairs with a single neural model (Johnson et al., 2017). MNMT also showed a promising emergent ability to translate in zero-shot directions between language pairs for which no parallel data was provided during training (Haddow et al., 2022). However, prior works mainly focused on training with enormous parallel data to prevent the degeneration in learning linguistic diversity from numerous languages (Dabre et al., 2020), while it remains unclear how it works for low-resource languages (LRLs). For example, more than 700 languages are spoken in Indonesia, where most are LRL or extreme-LRL, causing failure in generalisation (Aji et al., 2022; Bang et al., 2023).

Currently, utilising the prior knowledge of a pre-trained model is reported to benefit the LRLs (Haddow et al., 2022; Wang et al., 2021). Assuming the availability of pre-trained multilingual language models, we aspire to find effective methods to integrate new unseen languages. However, leveraging a pre-trained model is difficult due to the intricacy of extracting linguistic knowledge from higher resources. Even with observed languages, the zero-shot translation direction has suffered from failures in generating text with correct languages (Gu et al., 2019). We hypothesised that one culprit is the entanglement in linguistic representation that creates a bias towards higher-resource languages.

To resolve this problem, we employed two methods that regularise multilingual representation. The first method adds an auxiliary objective to predict the target language (Yang et al., 2021), regularising the representation in the decoder such that the tokens are geared towards the target language. The second method separates some layers in the decoder into shared and language-specific (LS) (Qu and Watanabe, 2022), effectively loosening unintended connections among language pairs and preserving more consistent and universal linguistic information, such as semantics.

We explored the effectiveness of the employed MNMT methods for Indonesia's LRLs, focusing on five languages: Indonesian, Javanese, Sundanese, Balinese, and Madurese, extracted from a subset of NusaX dataset (Winata et al., 2023). The dataset is an instance of the LRL dataset comprising a parallel corpus of ten of Indonesia's indigenous languages consisting of 1,000 samples for each language, and we used only half for the train set. Empirical results showed BLEU improvement as effectively as 7.4 at best, from the baseline score of 7.1 to 15.5 using the LS-layer separation.

Detail analysis indicates that the representation of the encoder gets disentangled by splitting the LS representation in the decoder. The disentanglement effectively mitigated the bias towards higher-resource languages, as reflected by a finer Off Target rate for the Indonesian. We also observed the superior transfer ability for languages unseen during pre-training, justified by the improvements in BLEU w.r.t. seen languages, and this revelation offers a promising insight for translating LRLs, i.e.

LREC-COLING 2024, pages 4978–4989

Indonesian languages.

2. Methodologies

2.1. Multilingual NMT

Given a pre-trained multilingual model on m languages, we trained a Multilingual Neural Machine Translation (MNMT) model that supports $n > m$ languages $\mathbb{L} = \{l_1, l_2, \ldots, l_n\}$ by fine-tuning the pre-trained model in which $l_{m+1}, \ldots, l_n$ are unseen during the pre-training. The pre-training maximised the probability of predicting the original input sequence with a noising function.

We employed a multilingual language model IndoBART (Cahyawijaya et al., 2021) as our pre-trained model, which was trained in the same manner as done by mBART (Liu et al., 2020) on Indonesian (`ind`) and two indigenous languages: Javanese (`jav`) and Sundanese (`sun`). When fine-tuning IndoBART, we re-used the tokeniser, the token embeddings, and the language tags (LT). For new unseen languages, i.e. Balinese (`ban`) and Madurese (`mad`), LTs were initialised from `ind`.

When tuning for the MNMT task, we followed Johnson et al. (2017) to ignore source LT and placed target LT (l) on the encoder side. Wu et al. (2021) found that the aforementioned approach captures more consistent semantic representations and makes the decoder attend more to the target language, which aligned with our preliminary experiment[1]. Given an input token sequence $\boldsymbol{x} \in \mathbb{R}^{|\boldsymbol{x}|}$, the model maximises the probability of outputting token sequence $\boldsymbol{y} \in \mathbb{R}^{|\boldsymbol{y}|}$ from target language $l \in \mathbb{L}$ by minimising the cross-entropy loss over the training data $\mathcal{T}$.

$$L_{MNMT} = -\sum_{(\boldsymbol{x},\boldsymbol{y},l)\in\mathcal{T}} \log p(\boldsymbol{y}|\boldsymbol{x}, l) \tag{1}$$

2.2. TLP: Target Language Prediction

During fine-tuning, we followed Yang et al. (2021) to employ an extra objective of language identification prediction. Specifically, given the decoder's last hidden state as an input, two Transformer encoder layers map the input into a representation $\mathbf{H} = [\boldsymbol{h}_1, \ldots, h_{|\boldsymbol{y}|}]$, where $\mathbf{H} \in \mathbb{R}^{|\boldsymbol{y}|\times d}$, which is then used to derive an average-pooled representation $\bar{\boldsymbol{h}} = \frac{\sum_{i=1}^{|\boldsymbol{y}|} \boldsymbol{h}_i}{|\boldsymbol{y}|}$. A classifier uses the pooled representation to predict the target language and obtains a cross-entropy loss, L_{TLP}.

$$L_{TLP} = -\sum_{\mathcal{T}} \log p(l|\bar{\boldsymbol{h}}) \tag{2}$$

[1] Details can be found in Appendix F

Training is performed by minimising the weighted[2] linear combination of L_{MNMT} and L_{TLP}.

$$L = (1-\alpha)\cdot L_{MNMT} + \alpha \cdot L_{TLP} \tag{3}$$

2.3. CLL: Central-Language-aware Layer

As an alternative method for TLP, we followed Qu and Watanabe (2022) that separates non-central language-specific (LS) information on the decoder side, by adding non-shared LS-layers (*LSL*) for languages other than `ind` that share the structure of the Transformer Vaswani et al. (2017) decoder's Feed-Forward Network (*FFN*). More concretely, let us denote the last layer output of encoder as $\mathbf{E} \in \mathbb{R}^{|\boldsymbol{x}|\times d}$, the k^{th} layer input of decoder as $\mathbf{H}_k \in \mathbb{R}^{|\boldsymbol{y}|\times d}$, and the target language as l, with *MHA* stands for MultiHeadAttention for a query and key/value pair (Vaswani et al., 2017) and *LN* for Layer Normalisation (Ba et al., 2016) then the Transformer decoder for CLL is defined as follows.

$$\mathbf{H}'_k = \mathit{LN}(\mathit{MHA}(\mathbf{H}_k, \mathbf{H}_k) + \mathbf{H}_k) \tag{4}$$
$$\mathbf{H}''_k = \mathit{LN}(\mathit{MHA}(\mathbf{H}'_k, \mathbf{E}) + \mathbf{H}'_k) \tag{5}$$
$$\mathbf{H}_{k+1} = \mathit{LN}(\mathit{CLL}(\mathbf{H}''_k, l) + \mathbf{H}''_k) \tag{6}$$

With w_l as the weight coefficient for l, the definition of *CLL* layer is as follows.

$$\mathit{CLL}(\mathbf{H}'', l) = \begin{cases} \mathit{FFN}(\mathbf{H}'') & \text{if } l = \texttt{ind}, \\ \mathit{FFN}(\mathbf{H}'') + w_l \mathit{LSL}_l(\mathbf{H}''). \end{cases} \tag{7}$$

3. Experiments

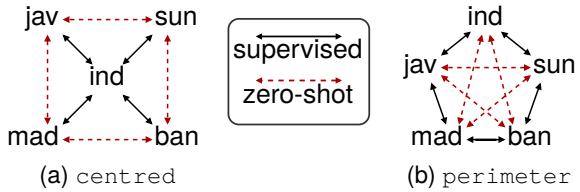
3.1. Setup

We employed Hugging Face Transformers (Wolf et al., 2020) as a framework for our experiments. We evaluated our experiments on the NusaX dataset (Winata et al., 2023). Following the default split of 5:1:4 for the train:valid:test, respectively, there are effectively only 500 samples per direction for training[3]. We chose only a subset of five languages: three languages used in pre-training of IndoBART, i.e. `ind`, `jav`, and `sun`, and two other languages not included in IndoBART, i.e. `ban` and `mad`, to represent the case of extending the model to new languages. We obtained the dataset through NusaCrowd (Cahyawijaya et al., 2023).

Zero-shot We introduce two settings for the zero-shot scenario. The first one is `centred` setup (Figure 1a), a commonly used setting in MNMT task, in which `ind` is chosen as central language and treat

[2] We set $\alpha = 0.1$ by following Yang et al. (2021).

[3] The details regarding hyper-parameters and actual dataset splits are in Appendix B & C, respectively.



(a) centred (b) perimeter

Figure 1: Zero-Shot setups.

Metric	Model	Centred		Perimeter	
		sup.	*z.s.*	*sup.*	*z.s.*
BLEU (↑)	Base	33.77	6.20	25.75	14.23
	TLP	34.58	6.21	26.07	15.02
	CLL	**35.06**	6.68	**27.42**	**22.24**
	CLL+TLP	34.78	**6.72**	27.23	22.18
% Off Target (↓)	Base	7.44	98.40	14.88	53.12
	TLP	3.94	97.79	13.93	50.83
	CLL	**1.06**	**94.08**	2.05	4.75
	CLL+TLP	1.47	94.38	**1.97**	**4.40**

Table 1: Average result by dataset setups with the number in bold denotes best performance. We refer to supervised as *sup.* and zero-shot as *z.s.*

all pairs with this language as supervised directions and the remaining as zero-shot. The second one is perimeter setup (Figure 1b), a non-centric setup to analyse the effect of disentanglement, in which each language is directly supervised in two directions, leaving the other two as zero-shot.

Evaluation Metrics We employed two metrics to measure performance. Firstly, *BLEU* score computed using SacreBLEU (Post, 2018). Secondly, *Off Target* computes the false negative ratio of the expected target languages. To our knowledge, no open off-the-shelf language detector supports all five languages, particularly ban and mad. As such, we trained a language detector module using FastText (Joulin et al., 2016) and described it further in Appendix A.

3.2. Results

Generally, CLL and CLL+TLP models are better in all directions measured by average BLEU in Table 1. Both have the average BLEU scores improved, with zero-shot directions having a better impact on the Off Target ratio. To be precise, in zero-shot, CLL improvement w.r.t. baseline is ∼.5 higher BLEU score with ∼4% finer Off Target in centred, and ∼8.0 higher BLEU score with ∼48% finer Off Target in perimeter counterpart. As for the TLP model, it has a slight improvement overall. For centred, we observed a better impact in the supervised direction with ∼.8 BLEU, and for perimeter, the better impact is in zero-shot direction with ∼.8 BLEU.

The translation direction jav→∗ in perimeter (Table 2) showed a complete combination of language pairs, that is, one zero-shot and supervised direction each from the seen and unseen target languages. The performance improvement w.r.t. baseline reflects the CLL model capability of adapting to unseen languages, i.e. in the supervised direction of jav→mad, we observed a higher ∼2.6 BLEU score with a finer Off Target ratio from 19.5% down to 0.75%. Similarly, a considerable improvement in the zero-shot direction of jav→ban is observed, with a ∼7.4 higher BLEU score and finer Off Target ratio from 96.75% down to 11.25%. In the TLP model, both supervised directions have better Off Target ratios but do not yield better BLEU scores, suggesting no direct correlation between predicting target language and better text generation.

Metric	Model	Supervised		Zero-shot	
		→ind	→***mad***	→sun	→***ban***
BLEU (↑)	Base	45.97	17.13	25.03	7.10
	TLP	45.35	17.32	25.97	7.17
	CLL	**47.42**	**19.73**	25.76	**15.53**
	CLL+TLP	47.42	19.69	**26.65**	14.78
% Off Target (↓)	Base	11.75	19.50	4.25	96.75
	TLP	13.00	19.00	2.00	96.25
	CLL	2.50	**0.75**	2.50	**11.25**
	CLL+TLP	**2.25**	1.50	**1.75**	11.50

Table 2: Result of jav→∗ in perimeter with the number in bold denotes best performance. Unseen languages during pre-training are in bold italics.

Model	ind →∗	Predicted Language ind	jav	sun	***ban***	***mad***	% Off Target	BLEU
Base	jav	0	397	1	2	0	0.75	40.07
	sun	2	0	398	0	0	0.50	37.45
	ban	9	7	2	228	154	**43.00**	**17.12**
	mad	11	2	7	34	346	**13.50**	**23.77**
CLL	jav	0	397	1	2	0	0.75	39.36
	sun	2	1	397	0	0	0.75	37.83
	ban	5	6	2	387	0	**3.25**	**24.92**
	mad	3	3	2	0	392	**2.00**	**26.20**

Table 3: Target-language confusion matrix of supervised directions ind→ ∗. Each has 400 cases. Unseen languages in pre-training are in bold italics.

The improvement of the CLL model is attributed mainly to the translation directions involving unseen languages as the target[4], as partially reflected from the Table 2. The result also shows the struggle for the baseline and TLP model in the directions involving unseen languages, as reflected by the bigger gap w.r.t. seen languages. And thus, we analysed and discussed the case further.

Sample cases are under Appendix E.

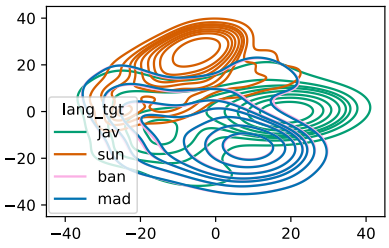
4. Analysis and Discussion

The improvement of the CLL model in terms of overall BLEU score is closely related to the finer Off Target ratio, which is the ability to generate text in the correct language. In the supervised

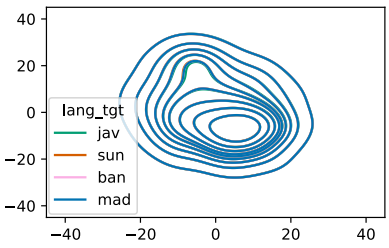
[4]Complete result is appended in Appendix D

Model	jav→∗ (*sup.*?)	Predicted Language jav	ind	sun	***ban***	***mad***	% Off Target	BLEU
Base	ind (✓)	6	353	0	9	32	11.75	46.0
	sun (✗)	4	13	383	0	0	4.25	25.0
	ban (✗)	3	232	0	13	152	**96.75**	**7.1**
	mad (✓)	0	78	0	0	322	**19.50**	**17.1**
CLL	ind (✓)	6	390	1	2	1	2.50	47.4
	sun (✗)	4	6	390	0	0	2.50	25.8
	ban (✗)	25	16	2	355	2	**11.25**	**15.5**
	mad (✓)	0	3	0	0	397	**0.75**	**19.7**

Table 4: Target-language confusion matrix of case jav→ ∗ in perimeter. Unseen languages during pre-training are in bold italics.



(a) Base



(b) CLL

Figure 2: Encoder representation for ind→∗.

case of ind→ ∗ (Table 3), for models trained in centred setup, the baseline has high counts in confusion (highlighted in red) between languages unseen during pre-training, i.e. ban and mad. On the other hand, the CLL model has zero Off Target count between the two languages and led to a better performance justified by the evaluation metrics (highlighted in blue).

The phenomenon is much more evident in the perimeter setup. In the case of jav→∗ as shown in Table 4, which has the same direction as Table 2, the baseline has the highest confusion to ind (highlighted in red), which has the highest magnitude of data during pre-training. As reflected in the zero-shot to unseen language (ban), for 96% of the time, it is off towards supervised direction 232 times and 152 times to ind and mad, respectively. The CLL model improves the situation remarkably (highlighted in blue) by refining the Off Target ratio for more than 85%. However, the CLL's Off Target comes from untranslated text or translated back to jav, a different tendency than the baseline.

Furthermore, we visualised the encoder representation on 2D space using t-SNE and bivariate KDE (Figure 2) with the case from the same direction as Table 3: centred-setup, ind→ ∗. The baseline model (Figure 2a) shows heavy entanglement from the three centroids: one each for jav and sun, and a shared centroid for ban and mad that is intertwining more towards jav. On the contrary, the CLL model (Figure 2b) has only one centroid invariant to the target language. This observation suggests the baseline has started translation to a specific target during encoding, while the CLL has the encoder representation separated from a target language-specific information, such as syntax and morpheme. The perimeter setup also indicates similar trends and defers the details to Appendix G.

Setup	Model	Perplexity (↓) jav→	***mad***→
centred	Base	1.232	**1.269**
	CLL	1.249	1.287
perimeter	Base	1.281	1.365
	CLL	**1.207**	1.340

Table 5: Averaged perplexity for ∗→ind on spelling style variations. Unseen languages in bold italics.

Analysis for each layer were also conducted by incorporating SVCCA (Raghu et al., 2017), which involves a Singular Vector Decomposition and is followed by a Canonical Correlation Analysis. The result supported our observation in the previous paragraph, and defers the details to Appendix H.

Additionally, we investigated the impact of naturally observed spelling style variations in digital texts as most of Indonesia's LRLs are colloquial, used alongside ind, and no institution standardises the spellings[5]. Table 5 shows the case of ∗→ind for two languages, jav and mad, representing seen and unseen language, respectively. In the case of the seen language of jav, where several spelling variations occurred in the pre-training data, the CLL model trained in perimeter setup has the lowest perplexity, suggesting a more consistent representation. In the case of the unseen language of mad, lower perplexity is observed for models trained under centred setup, indicating a bias towards ind that might be coming from the spurious correlation.

5. Related Works

In the LRL context, Lee et al. (2022) pointed out the inability of multilingual models to adapt to languages unseen during pre-training, i.e. mBART and mT5. However, they did not consider separating the LS layers during tuning steps nor incorporat-

[5]see samples in Appendix I

ing representation disentanglement methods. On the other hand, works that put some consideration on this did not consider the usage in LRL via transfer learning, such as shown by Philip et al. (2020). Complementing the prior works, we showed how BART architecture with LS layers worked on LRL even when applied only during the fine-tuning step, effectively leveraging prior knowledge to adapt to unseen languages.

6. Conclusion

Representation entanglement is constraining the capabilities of the MNMT system. With the empirical results on Indonesia's Austronesian LRLs, we revealed how the disentanglement of multilingual representation helped de-biasing from higher resources and boosted the performance of the LRL system even in zero-shot conditions and on new languages unseen during pre-training. This revelation suggests a promising direction for adopting LRLs into the existing system.

7. Limitations

In this work, we did not consider the integration of new characters and assumed the sufficiency of a pre-trained model. We made the assumption because most of Indonesia's LRLs are used alongside the official Indonesian language and are transliterated using 26 Latin characters from the Indonesian alphabet. It is also worth noting that our empirical results are from the languages within the same Malayo-Polynesian subgroup of the Austronesian language family, and the extent to which this affects is unclear yet.

8. Bibliographical References

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Muhammad Satrio Wicaksono, Ivan Parmonangan, Ika Alfina, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Septiandri, James Jaya, Kaustubh Dhole, Arie Suryani, Rifki Afina Putri, Dan Su, Keith Stevens, Made Nindyatama Nityasya, Muhammad Adilazuarda, Ryan Hadiwijaya, Ryandito Diandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Damapuspita, Haryo Wibowo, Cuk Tho, Ichwanul Karo Karo, Tirana Fatyanosa, Ziwei Ji, Graham Neubig, Timothy Baldwin, Sebastian Ruder, Pascale Fung, Herry Sujaini, Sakriani Sakti, and Ayu Purwarianti. 2023. NusaCrowd: Open source initiative for Indonesian NLP resources. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13745–13818, Toronto, Canada. Association for Computational Linguistics.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, Ayu Purwarianti, and Pascale Fung. 2021. IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

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

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

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.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

En-Shiun Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Adelani, Ruisi Su, and Arya McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 58–67, Dublin, Ireland. 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.

Jerin Philip, Alexandre Berard, Matthias Gallé, and Laurent Besacier. 2020. Monolingual adapters for zero-shot neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4465–4470, Online. Association for Computational Linguistics.

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

Zhi Qu and Taro Watanabe. 2022. Adapting to non-centered languages for zero-shot multilingual translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5251–5265, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6076–6085. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Rui Wang, Xu Tan, Renqian Luo, Tao Qin, and Tie-Yan Liu. 2021. A survey on low-resource neural machine translation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4636–4643. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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

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

9. Language Resource References

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Appendices

A. Language Detector for Off Target Evaluation

FastText library (Joulin et al., 2016) was employed to train language detector module on the data-set defined in section 3.1. This module has an objective to classify text into the *five languages* experimented in this work. Accuracy performance of this language detector module is as listed on table 6.

Data Split	Sample Count	Accuracy
train	2,500	100%
valid	500	98.8%
test	2,000	98.9%

Table 6: Language Detector Performance

The hyper-parameter settings are left to default unless specified: `'dim=12'`, `'lr=0.088'`, `'epoch=75'`, `'wordNgrams=1'`.

B. Model Hyper-parameters

Hyper-parameters during training is as defined in Table 7. The weight coefficient for TLP Loss, $alpha$, is set at 0.1 following the original work by Yang et al. (2021).

Key	`centred`	`perimeter`
early-stop patience	12 evals	15 evals
evaluation per epoch	4	5
steps per epoch	200	250
max. epoch	20 epochs	
batch-size	20 sentence-pairs	
optimiser	Adam	
learning-rate	1×10^{-5}	
weight decay	0.01	

Table 7: Training-time hyper-parameters

Additional hyper-parameters during test-time are left default unless specified: `'maxlen=128'`, `'top_k=4'`, `'top_p=0.95'`.

C. Dataset Split Details

NusaX dataset (Winata et al., 2023) is a parallel corpus that consists of 1,000 sentences for each language. There are 10 indigenous languages of Austronesian family plus Indonesian (`ind`) and English. Default split ratio for train : valid : test of $5:1:4$ was applied. Actual data-split used in this experiment is listed in table 8. Corresponding count during training-time in `centred`-setup and `perimeter`-setup are listed in table 9 and table 10, respectively.

Data Split	`centred`		`perimeter`	
	sup.	z.s.	sup.	z.s.
train	4,000	-	5,000	-
valid	800	-	1,000	-
test	3,200	4,800	4,000	4,000

Table 8: Dataset split

Direction	`jav`	`ind`	`sun`	`ban`	`mad`
`jav`→∗	-	500	0	0	0
`ind`→∗	500	-	500	500	500
`sun`→∗	0	500	-	0	0
`ban`→∗	0	500	0	-	0
`mad`→∗	0	500	0	0	-

Table 9: Training dataset in `centred`-setup

Direction	`jav`	`ind`	`sun`	`ban`	`mad`
`jav`→∗	-	500	0	0	500
`ind`→∗	500	-	500	0	0
`sun`→∗	0	500	-	500	0
`ban`→∗	0	0	500	-	500
`mad`→∗	500	0	0	500	-

Table 10: Training dataset in `perimeter`-setup

D. Evaluation Result Details

Metric	Model	ind2sun	ind2ban	ind2mad	ind2jav	sun2ind	ban2ind	mad2ind	jav2ind
BLEU (↑)	Base	37.45	17.12	23.77	40.07	**41.82**	**35.39**	29.61	44.92
	TLP	37.65	21.76	25.48	**40.56**	41.63	34.83	29.71	45.00
	CLL	**37.83**	**24.92**	**26.20**	39.36	41.82	35.24	**29.81**	**45.34**
	CLL+TLP	37.60	24.03	25.41	39.98	41.74	35.12	29.37	44.97
% Off Target (↓)	Base	**0.50**	43.00	13.50	**0.75**	**1.00**	**0.25**	0.25	**0.25**
	TLP	0.75	22.25	6.25	**0.75**	**1.00**	**0.25**	**0.00**	**0.25**
	CLL	0.75	**3.25**	**2.00**	**0.75**	1.25	**0.25**	**0.00**	**0.25**
	CLL+TLP	**0.50**	4.00	2.50	1.75	1.25	1.25	**0.00**	0.50

Table 11: **`centred` supervised** direction.

Metric	Model	sun2ban	sun2mad	sun2jav	ban2sun	ban2mad	ban2jav	mad2sun	mad2ban	mad2jav	jav2sun	jav2ban	jav2mad
BLEU (↑)	Base	7.61	4.18	6.91	7.35	4.38	6.10	6.12	6.23	5.09	8.30	7.69	4.49
	TLP	7.54	4.16	6.92	7.28	4.30	6.33	6.12	6.08	5.14	8.46	7.84	4.41
	CLL	7.86	5.13	7.63	**7.61**	**4.81**	7.38	**6.26**	**6.27**	5.68	**8.52**	**8.23**	4.75
	CLL+TLP	**8.02**	**5.19**	**7.92**	7.56	4.74	**7.46**	6.23	6.11	**5.96**	8.48	8.17	**4.80**
% Off Target (↓)	Base	100.00	100.00	96.75	97.75	99.75	96.25	95.75	100.00	97.50	97.25	100.00	99.75
	TLP	100.00	99.75	95.75	97.25	99.50	95.75	**92.50**	99.75	96.75	96.75	100.00	99.75
	CLL	98.75	**92.25**	91.25	96.25	**92.75**	**88.75**	**92.50**	98.25	90.00	**96.25**	**98.00**	**94.00**
	CLL+TLP	**97.50**	93.00	**90.75**	**95.75**	**92.75**	92.00	94.25	**98.00**	**88.00**	**96.25**	99.25	95.00

Table 12: **`centred` *zero-shot*** direction.

Metric	Model	ind2sun	ind2jav	sun2ind	sun2ban	ban2sun	ban2mad	mad2ban	mad2jav	jav2ind	jav2mad
BLEU (↑)	Base	34.64	38.13	36.58	11.24	23.34	16.53	12.40	21.54	45.97	17.13
	TLP	**34.83**	**38.71**	36.51	11.81	**23.64**	**17.28**	12.90	**22.32**	45.35	17.32
	CLL	34.15	36.78	43.93	**16.55**	23.35	16.66	**14.03**	21.62	**47.42**	**19.73**
	CLL+TLP	34.71	36.98	**44.09**	16.23	23.11	15.81	12.91	21.38	47.42	19.69
% Off Target (↓)	Base	**0.50**	2.50	27.25	64.00	**1.00**	5.25	16.00	**1.00**	11.75	19.50
	TLP	**0.50**	**1.50**	26.25	57.75	1.25	5.00	14.00	**1.00**	13.00	19.00
	CLL	**0.50**	1.75	3.00	7.00	1.25	0.75	**1.50**	1.50	2.50	**0.75**
	CLL+TLP	0.75	1.75	**2.75**	**6.50**	**1.00**	**0.50**	1.75	**1.00**	**2.25**	1.50

Table 13: **`perimeter` supervised** direction.

Metric	Model	ind2ban	ind2mad	sun2mad	sun2jav	ban2ind	ban2jav	mad2ind	mad2sun	jav2sun	jav2ban
BLEU (↑)	Base	11.33	12.60	4.89	21.56	15.57	15.43	10.48	18.30	25.03	7.10
	TLP	11.03	13.09	5.25	24.24	16.24	17.13	11.28	18.78	25.97	7.17
	CLL	**18.72**	**19.97**	15.77	26.24	**33.00**	21.07	**27.71**	18.57	25.76	**15.53**
	CLL+TLP	17.58	19.71	**16.10**	**26.33**	32.99	**21.23**	27.51	**18.94**	**26.65**	14.78
% Off Target (↓)	Base	86.50	55.50	89.00	15.50	69.50	33.25	80.75	0.25	4.25	96.75
	TLP	80.75	51.50	88.50	10.00	69.25	31.25	78.75	**0.00**	2.00	96.25
	CLL	8.75	6.50	3.50	**2.25**	3.50	**3.75**	5.00	0.50	2.50	**11.25**
	CLL+TLP	**7.75**	**5.75**	**2.75**	2.75	**3.25**	4.00	**4.25**	0.25	**1.75**	11.50

Table 14: **`perimeter` *zero-shot*** direction.

In the `centred`-setup, TLP and CLL improved the supervised direction when translating to unseen languages, i.e.: `ind`→{`ban`,`mad`}, resulting in BLEU scores closer to seen languages. In the zero-shot direction, improvement of TLP and CLL is minuscule.

The effect of TLP and CLL could be observed clearer in the `perimeter`-setup. On the top of improving the scores of unseen languages like the `centred`-setup, the performances followed the amount of data fed into the model including during pre-training, i.e. central (`ind`) > seen (`jav`, `sun`) > unseen (`ban`, `mad`). The pre-training dataset comprises 98.5% `ind` and 1.5% combination of `jav` and `sun`.

E. Example Testcases

Reference (ind)		mayoritas pengguna hak pilih di pemilihan gubernur telah menentukan pilihan.
Source (jav)		mayoritas sing nganggo hak pilih ing pemilihan gubernur uwis nentukake pilian.
Gold (ban)		mayoritas sane nganggen hak pilih ring pemilihan gubernur sampun nentuang pilihan.
centred	Base	mayoritas tidak memiliki hak pilih di pemilihan gubernur sudah memilih pilian. (ind)
	CLL	mayoritas yang menggunakan hak pilih di pemilihan gubernur telah memilih pilian. (ind)
perimeter	Base	mayoritas yang menggunakan hak pilih di pemilihan gubernur sudah menghasilkan pilian. (ind)
	CLL	mayoritas sane nganggo hak pilih ring pemilihan gubernur nenten nyajiang pilian. (ban)

Reference (ind)		restonya bagus dan besar. waktu ke sini saat makan siang dan hari minggu ramai sekali pengunjung. makanannya enak enak, gurame goreng dan sambalnya mantap.
Source (jav)		restone apik lan gede. wektu marang rene pas mangan siang lan dina minggu rame tenan pengunjung. panganane enak enak, gurame goreng lan sambele mantep.
Gold (ban)		restone luung lan gede. dugase mai pas ngajeng tengai lan hari minggu rame sajan pengunjung. ajengane jaan-jaan, gurame goreng lan sambalne mantap.
centred	Base	restenya bagus dan gede. waktu ke sini pas makan siang dan hari minggu rame sekali pengunjung. makanannya enak enak, gurame goreng dan sambal mantep. (ind)
	CLL	restinya mantap dan gede. waktu ke sini pas makan siang dan hari minggu rame sekali pengunjung. makanannya enak enak, gurame goreng dan sambal mantep. (ind)
perimeter	Base	restona bhegus ben bhegal. bekto ka diye pas ngakan siang ben malem are rame sarah pengunjung. kakananna nyaman nyaman, gurame goreng ben sambele mantep. (mad)
	CLL	restone becik lan gede. waktu mai pas ngajeng tengai lan dina minggu rame sajan pengunjung. ajengane jaen jaen, gurame goreng lan sambalne mantap. (ban)

Table 15: De-tokenised text of zero-shot jav→ban case with a lot of non-translate-able Indonesian terminologies. CLL trained under perimeter showed unerring translation to the target language, while others biased towards highest-resources ind, except for perimeter-setup base model in bottom-case which generated text of mad.

F. Language Tag Matters: A Preliminary Experiment

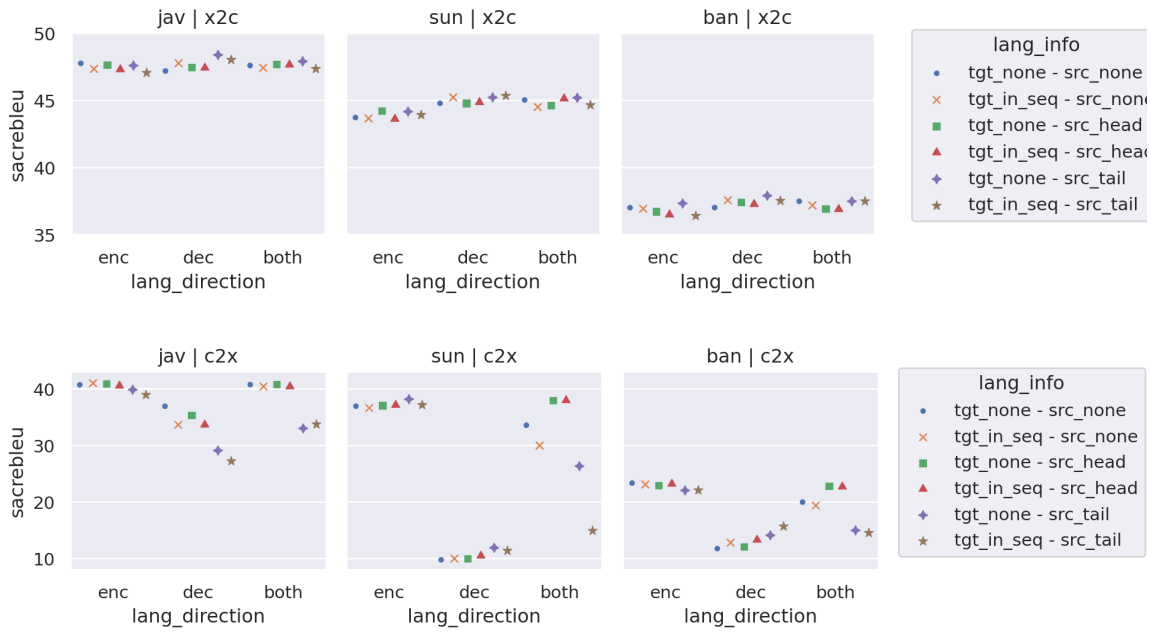


Figure 3: Various Language-Tag (LT) configuration performances. lang_direction represents where the target-LT is located, while source-LT is denoted by src_*. No source-LT is provided for src_none, otherwise it is located in the encoder side.

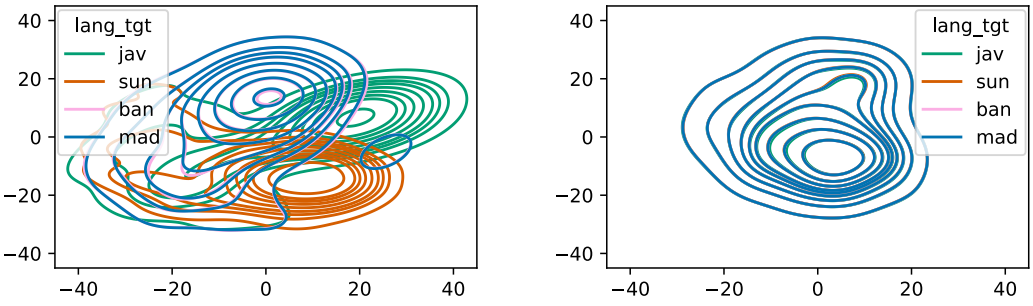
A preliminary experiment to check the effectiveness of Language Tag configuration was conducted. We

followed a portion of the experiments from Wu et al. (2021). Please refer to the original work for a more thorough and clearer experimental details.

G. Encoder Representation Visualisation

G.1. `perimeter`-setup, `ind`→∗

A comparison to Figure 2 with direction `ind`→∗ but trained under `perimeter` setup, showing similar phenomenon as explained in Section 4.

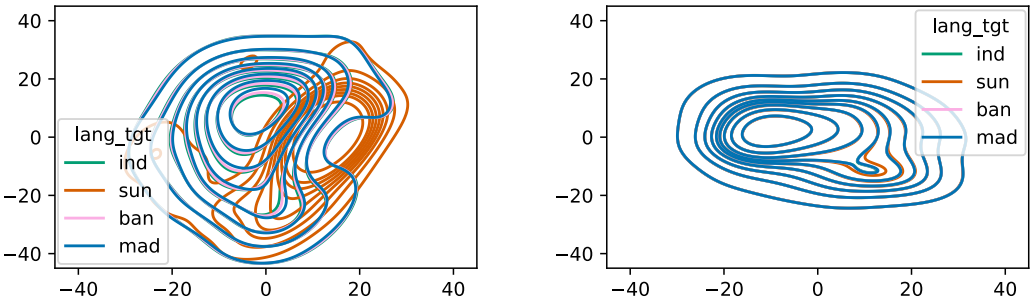


(a) Baseline (b) CLL

Figure 4: t-SNE visualisation of `ind`→∗ in `perimeter`.

G.2. `perimeter`-setup, `jav`→∗

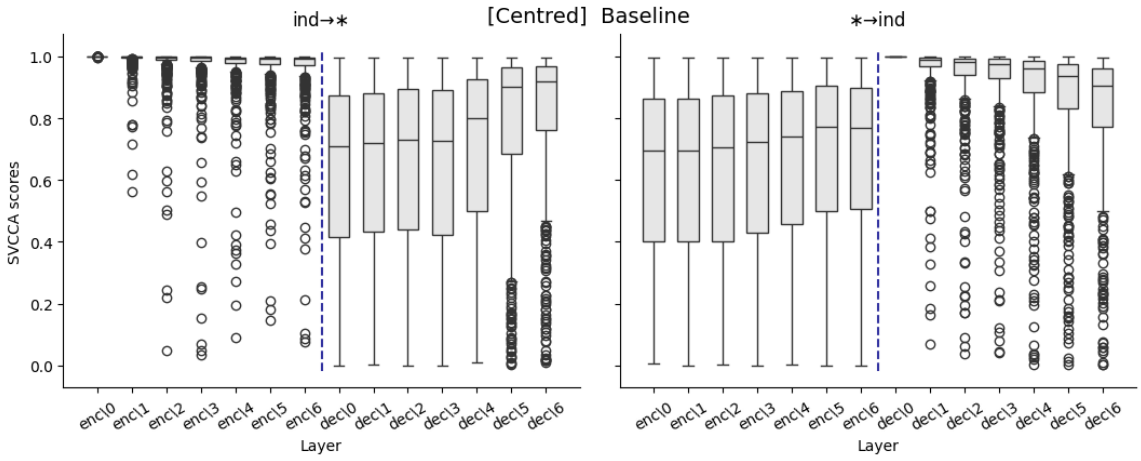
Visualisation of translation direction from `jav`→∗, the same direction as Table 2 & 4, where exists one direction from each combination of supervision type and seen-unseen languages during pre-training.



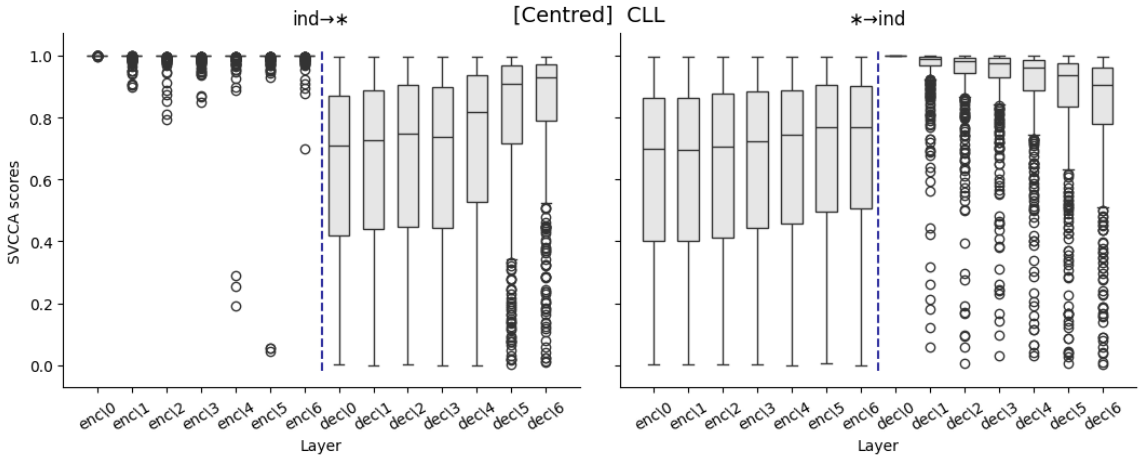
(a) Baseline (b) CLL

Figure 5: t-SNE visualisation of `jav`→∗ in `perimeter`.

H. Per Layer Correlation



(a) Baseline



(b) CLL

Figure 6: Distribution plot of SVCCA for `centred`-setup.

I. Text Variations Case

In jav, certain words have the same phoneme but are spelt with either 'b' or 'p', whereas in mad, sometimes it is not compulsory to spell 'h' and 'y'.

Direction	Input Text (raw)	Perplexity			
		centred		centred	
		Base	CLL	Base	CLL
jav→ind	Restone apik lan gede. Wektu marang rene pas mangan siang lan dina minggu rame tenan pengunjung. Panganane enak enak, gurame goreng lan sambele mantep.	1.1807	1.2168	1.1807	**1.1579**
	Restone apik lan gede. Wektu mrene pas mangan awan lan dina minggu rame tenan pengunjung. Panganane enak enak, gurame goreng tur sambele mantep.	1.2497	1.2601	1.3858	**1.2044**
	Restone apik lan gede. Wektu marang rene pas mangan siang lan dina minggu rame banget pengunjunge. Panganane enak enak, gurame goreng tur sambele manteb.	**1.1988**	1.2476	1.2656	1.2132
	Restone apik lan gede. Wektu mrene pas mangan awan lan dina minggu rame banget pengunjunge. Panganane enak enak, gurame goreng tur sambele manteb.	1.2677	1.2781	1.3314	**1.2567**
mad→ind	Restona bhegus ben raje. Bekto ka diye teppa'na ngakan aben ben are minggu rammi sarah se deteng. Kakananna man-nyaman, gurame ghuring ben sambhelle mantap.	**1.2608**	1.2641	1.3426	1.2953
	Restona bhegus ben raje. Bekto ka diye pas ngakan aben ben are minggu cek rammena se deteng. Kakananna man-nyaman, gurame ghuring ben sambel la nyaman.	1.2707	**1.2636**	1.3704	1.3269
	Restona begus ben raje. Bekto ka die teppa'na ngakan aben ben are minggu rammi sarah se deteng. Kakananna man nyaman, gurame guring ben sambelle mantap.	**1.2777**	1.3298	1.3737	1.3171
	Restona begus ben raje. Bekto ka die pas ngakan aben ben are minggu cek rammena se deteng. Kakananna man nyaman, gurame guring ben sambel la nyaman.	**1.2792**	1.3285	1.4050	1.4006

Table 16: Example Cases of Text Variations.