IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 July 2023, accepted 22 August 2023, date of publication 25 August 2023, date of current version 30 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3308818

RESEARCH ARTICLE

Many-to-Many Multilingual Translation Model for Languages of Indonesia

WILSON WONGSO, ANANTO JOYOADIKUSUMO, BRANDON SCOTT BUANA, AND DERWIN SUHARTONO, (Member, IEEE)
Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia
Corresponding author: Wilson Wongso (wilson.wongso001@binus.ac.id)

ABSTRACT Indonesia is home to over 700 languages and most people speak their respective regional languages aside from the lingua franca. In this paper, we focus on the task of multilingual machine translation for 45 regional Indonesian languages and introduced Indo-T5 which leveraged the mT5 sequence-to-sequence language model as a baseline. Performances of bilingual and multilingual fine-tuning methods were also compared, in which we found that our models have outperformed current state-of-the-art translation models. We also investigate the use of religious texts from the Bible as an intermediate mid-resource translation domain for low-resource translation domain specialization. Our findings suggest that this two-step fine-tuning approach is highly effective in improving the quality of translations for low-resource text domains. Our results show an increase in SacreBLEU scores when evaluated on the low-resource NusaX dataset. We release our translation models for other researchers to leverage.

INDEX TERMS Languages of Indonesia, low-resource languages, mT5, natural language processing, neural machine translation.

I. INTRODUCTION

Indonesia is considered one of the most linguistically diverse countries in the entire world, estimated to host over 700 living languages [1]. This figure still excludes the additional regional dialects or minor lexical variations that exist in each unique language category located in different geographical locations [2]. Compounding the common occurrence of language mixing during colloquial conversations, the true magnitude of linguistic diversity in Indonesia becomes close to immeasurable. The majority of Indonesians today also continue to converse in regional languages instead of their primary lingua franca on a daily basis [3], [4].

Unfortunately, the number of existing translation models for Indonesia's language does not reflect the language diversity of the nation due to substantial challenges in developing NLP (natural language processing) systems for regional, underrepresented languages [1]. Recent works surrounding massively multilingual translation models like M2M-100 (Many-to-Many Multilingual Model) [5] and NLLB (No Language Left Behind) [6] have begun efforts towards under-resourced languages for the task of neural machine translation. These models were specifically trained and designed for machine translation and provided significant improvements in translation performance. Before translation-specific models like these, it is common to leverage generic sequence-to-sequence models such as T5 (Text-to-Text-Transfer-Transformer) [7] and BART (Bidirectional Auto-Regressive Transformers) [8]. Both of these models have multilingual variants called mT5 (Multilingual T5) [9] and mBART (Multilingual BART) [10], respectively.

These models try and cover as many languages as possible, with a wide variety of language families and subgroups. However, there are only a few languages of Indonesia that were included in them. Table 1 lists several languages of Indonesia that were covered in previous works. NLLB has the most number of languages of Indonesia within its corpus, while mBART only includes Indonesian (`ind`) and no regional languages at all. IndoBART [11], most recently, attempts to create an Indonesia-centric model by training on Indonesian (`ind`), Javanese (`jav`), and Sundanese (`sun`). These considerations encouraged us to create a many-to-many translation model with a focus on the languages of Indonesia.

The associate editor coordinating the review of this manuscript and approving it for publication was Sangsoon Lim.

TABLE 1. Languages of Indonesia covered in existent datasets and multilingual models.

ISO 639-3	Language	NusaX [12]	NLLB [6]	IndoBART [11]	mT5 [9]	M2M-100 [5]	mBART [10]
ace	Acehnese	✓	✓				
ban	Balinese	✓	✓				
bjn	Banjarese	✓	✓				
bug	Buginese	✓	✓				
ind	Indonesian	✓	✓	✓	✓	✓	✓
jav	Javanese	✓	✓	✓	✓	✓	
mad	Madurese	✓					
min	Minangkabau	✓	✓				
nij	Ngaju	✓					
sun	Sundanese	✓	✓	✓	✓	✓	
bbc	Toba Batak	✓					

We leveraged the pre-trained mT5 model [9] and further fine-tuned it as a translation model for the languages of Indonesia. Using religious texts as an intermediate, mid-resource text domain, we provided a strong model checkpoint from which end-users could further fine-tune to the specific text domain of their choice. To evaluate our proposed method, we conducted performance benchmarks on translation for social media texts [12]. Our models were then released on the HuggingFace Model Hub for other researchers to leverage.

II. RELATED WORKS

Neural machine translation methods have not only shown clear performance improvements over rule-based translation and statistical machine translation but also reduce the need to manually feature engineer the rules of translation given a pair of languages. For instance, [13] showed that contextualized, sequence-to-sequence language modeling is an effective way to train a bilingual neural machine translation model. Their Transformer has an encoder-decoder architecture and leverages the attention mechanism [14]. This allows the model to automatically learn soft translation alignments between the pair of languages during training.

Since its first advent, there has been a multitude of methods, improvements, and implementations based on the original Transformers architecture. For instance, GELU [15] was proposed to be used as an activation function, synthetic attention of Synthesizer variants [16] to replace self-attention, Switch Transformers [17] to scale the architecture's parameters efficiently, and still many others. In the realm of translation specifically, [18] proposed that back-translation, data filtering, and fine-tuning on domain-specific data, are methods to improve Transformer-based machine translation. Reference [19] similarly trained hundreds of Transformer-based machine translation models based on the efficient implementation of MarianMT [20]. Their machine translation models may either be bilingual or multilingual. The latter type is a single model that could translate to and from multiple languages.

Reference [21] scaled the idea of training a massively multilingual translation model and showed that they are not only more performant than their bilingual counterpart but also benefited low-resource settings. However, they used English as a pivot language, thus all languages are translated to and from English only. Reference [5] expanded this approach and further trained a truly multilingual model, M2M-100, that is able to translate between any of the 100 pairs of languages found in its training data. Recently, [6] escalated the technique to 200 pairs of languages, creating the current largest and reportedly most performant multilingual machine translation model.

GPT (Generative Pre-trained Transformer) [22] takes the decoder of Transformers, while BERT (Bidirectional Encoder Representations from Transformers) [23] takes the encoder, and both showed that unsupervised pre-training on abundant unlabelled text corpora provides a performant checkpoint to further fine-tune on downstream natural language understanding tasks. This idea of pre-training a language model on large text corpora was then applied to sequence-to-sequence models like T5 [7] and BART [8]. T5 converts a variety of language tasks into a text-to-text format, while BART learns to reconstruct the corrupted span of texts as their respective pre-training tasks. Both of them showed powerful results on downstream fine-tuning tasks like translation. Moreover, they both have multilingual variants called mT5 [9] and mBART [10], respectively. It has been shown that both mT5 and mBART can similarly be utilized for further fine-tuning as a translation model.

Pre-trained multilingual transformers are commonly used for low-resource languages such as languages of Africa [24], [25] and Urdu [26]. Despite boasting substantial speaker populations, these languages are impeded by limited online textual data, similar to that of the languages of Indonesia. Presently, only Indonesian, Javanese, and Sundanese are supported on Google Translate, accentuating the need to explore strategies to include low-resource languages in machine translation systems.

Over time, developments of language family-specific translation models have also begun to appear. For instance, the pre-trained checkpoint of M2M-100 [5] was leveraged to develop multiple bilingual machine translation models for

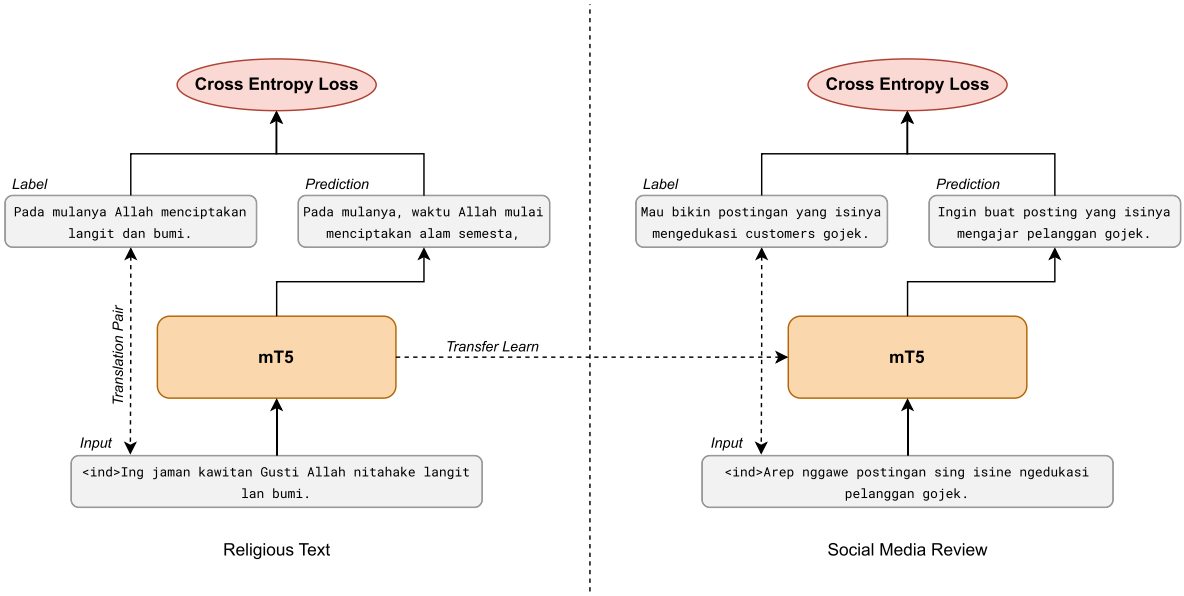


FIGURE 1. Proposed two-step machine translation model fine-tuning approach.

African languages, including transferability to a low-resource setting [24]. Likewise, mT5 [9] provided a strong checkpoint to be fine-tuned for a multilingual, many-to-many translation model for African languages [25]. In the context of the languages of Indonesia, IndoBART [11] is a pre-trained denoising model based on BART [8], trained on three languages (`ind`, `jav`, `sun`), and was shown to be an effective fine-tuning checkpoint for downstream sequence-to-sequence tasks [27].

Based on these previous works and considering the practicality of the desired outcome, we will be researching the efficacy of:

1) Designing a many-to-many, multilingual machine translation model for the languages of Indonesia.
2) Fine-tuning pre-trained checkpoints of multilingual sequence-to-sequence models such as mT5.
3) Examining the transferability of our translation model to low-resource text domains.

III. DATA METHODOLOGY

A. RELIGIOUS TEXTS AS INTERMEDIATE TEXT DOMAIN

Our main source of data for training are texts taken from the many translations of the Bible, sourced primarily from Alkitab Mobile SABDA.[1] Texts are aligned on a verse level, and they are assumed to be valid parallel translation pairs which will be learned by our model. From Alkitab Mobile SABDA alone, there are 45 languages of Indonesia,[2] as outlined in Table 2. However, the number of verses varies for every language. Some languages only have New Testament translations, containing around 7,900 verses. Nonetheless, certain languages do have both Old and New Testament translations, totaling up to around 31,000 verses. §VI-B investigated whether this discrepancy in the number of translation pairs affects the resultant translation model performance on the particular language.

To have an initial evaluation of our proposed method, we began with only a select number of languages instead of all of the languages laid out in Table 2, which we will refer to as `All`. We selected a subset of languages which are also supported by IndoBART and IndoGPT [11], NLLB [6], [29], [30], and NusaX [12], in order to have a fair comparison between these previous models with ours. This includes underlined languages found in Table 8 and we refer to this subset as `Baseline`.

This collection of Bible verses serves as mid-resource training data, with the aim of not only further fine-tuning the model for translating languages of Indonesia, but also introducing additional languages that were not present in the pre-training data of mT5 [9]. We summarized our overall approach in Fig. 1.

1) ALIGNING BIBLE VERSES

Bible verses in the different languages, however, are not partitioned identically, which causes issues when mapping verses between languages (or alignment), we categorize them into three types. The first cause of misalignment is *cascading verses* – additional verse identifiers which do not exist in another language. For instance, Exodus 6:1 in the Javanese Bible has an additional identifier "(5-24)", but this does not appear in the Balinese Bible. Similarly, Exodus 6:2 has the additional identifier "(6-1)" and so on. These extra verse identifiers, however, can simply be ignored, as the content of the verse will still be a valid translation pair between other

[1] https://alkitab.mobi/

[2] Some languages are traditionally written in non-Latin scripts, but their Romanized versions have become predominant in modern usages.

TABLE 2. List of languages in religious text data and their linguistic details. Data obtained from ethnologue [28]. (W.O. = word order).

ISO	Language	Primary Region	W.O.	#speakers
	Old Testament and New Testament			
ban	Balinese	Bali	SVO	3,300,000
bbc	Toba Batak	North Sumatra	N/A	1,610,000
bts	Batak Simalungun	North Sumatra	N/A	151,000
btx	Batak Karo	North Sumatra	SVO	491,000
bug	Buginese	South Sulawesi	SVO	4,370,000
ind	Indonesian	Indonesia	SVO	198,000,000
jav	Javanese	Java	SVO	68,200,000
mad	Madurese	Madura	SVO	7,790,000
mak	Makassarese	South Sulawesi	N/A	2,110,000
sda	Toraja-Sa'dan	South Sulawesi	N/A	588,000
sun	Sundanese	West Java	SVO	32,400,000
	New Testament only			
abs	Ambonese	Maluku	N/A	1,650,900
ace	Acehnese	Aceh	SVO	2,840,000
atq	Aralle-Tabulahan	South Sulawesi	N/A	29,300
bkl	Berik	Papua	SOV	1,200
blz	Balantak	East Sulawesi	N/A	20,500
btd	Pakpak Dairi	North Sumatra	N/A	172,000
bvz	Bauzi	Papua	SOV	1,500
gbi	Galela	North Maluku	N/A	79,000
gor	Gorontalo	Gorontalo	N/A	505,000
hvn	Hawu	Savu Island	N/A	110,000
iba	Iban	West Kalimantan	SVO	1,452,000
kgr	Abun	West Papua	N/A	3,000
kzf	Da'a Kaili	Central Sulawesi	N/A	62,600
ljp	Lampung Api	South Sumatra	N/A	827,000
mej	Meyah	West Papua	N/A	14,800
min	Minangkabau	West Sumatra	SVO	4,880,000
mkn	Kupang	West Timor	SVO	350,000
mog	Mongondow	North Sulawesi	N/A	117,000
mqj	Mamasa	West Sulawesi	N/A	89,100
mqy	Manggarai	Flores	N/A	900,000
mvp	Duri	Sulawesi	VSO	123,000
mwv	Mentawai	West Sumatra	N/A	62,300
nia	Nias	Nias	VOS	867,000
nij	Ngaju	Central Kalimantan	N/A	890,000
npy	Napu	Central Sulawesi	N/A	6,240
pmf	Pamona (Taa)	Central Sulawesi	N/A	77,900
ppk	Uma	Central, South Sulawesi	N/A	18,800
ptu	Bambam	West Sulawesi	N/A	42,100
sas	Sasak	Lombok	N/A	2,100,000
sxn	Sangirese	North Sulawesi	N/A	110,000
tby	Tabaru	North Halmahera	N/A	15,000
twu	Termanu	Rote Island	N/A	30,000
yli	Anggurk Yali	Papua	N/A	15,000
yva	Yawa	Papua	SOV	10,000

languages. This trend of cascading verses with additional verse identifiers can be found across different Biblical translations, as shown in Table 3.

In other scenarios, *multiple verses* in a specific language might also be combined, such as in Revelations 13:1. Fortunately, verse combinations across languages are parallel, i.e. Revelations 12:8 is combined with Revelations 13:1 in all of the languages. Therefore, we can again remove these verse identifiers and recognize the verse contents to be valid translation pairs. Table 4 displays an example of this trend.

TABLE 3. Exodus 6:1 in Javanese, Indonesian, and Balinese translations. Verse identifiers are highlighted in red. Dropping verse identifiers will not cause any alignment issues.

ISO	Verse Content
jav	(5-24) Nanging pangandikane Pangeran Yehuwah marang Nabi Musa: "Ing samengko sira bakal sumurup, apa kang bakal Suntandukake marang Pringon; anggone bakal nglilani lunga bangsa iki sarana dipeksa ing asta kang rosa, iya marga saka dipeksa dening asta kang rosa dheweke bakal nundhung bangsa iku saka ing nagarane."
ind	(5-24) Tetapi TUHAN berfirman kepada Musa: "Sekarang engkau akan melihat, apa yang akan Kulakukan kepada Firaun; sebab dipaksa oleh tangan yang kuat ia akan membiarkan mereka pergi, ya dipaksa oleh tangan yang kuat ia akan mengusir mereka dari negerinya."
ban	Ida Sang Hyang Widi Wasa raris ngandika ring Dane Musa: "Ane jani kita lakar nepukin saluiring ane lakar laksanayang Ulun marep teken sang prabu. Ulun lakar maksa ia, apang ia maang kaulan Ulune makaad. Sasajaane Ulun lakar maksa ia apanga ia nundung kaulan Ulune uli guminnyane."

TABLE 4. Revelations 13:1 in Javanese, Sundanese, and Madurese translations. Verse identifiers are highlighted in red. Dropping multiple verse identifiers will not cause any alignment issues.

ISO	Verse Content
jav	(12-18) lan banjur manggon ana ing pinggir sagara. (13-1) Sabanjure aku weruh ana kewan njedhul saka sajroning sagara, sungune sapuluh lan endhase pitu; pucuking sungune ana makuthane sapuluh, lan ing endhase ana cirine jeneng panyenyamah.
sun	(12-18) Gen naga ngajanteng di sisi basisir. (13-1) Ti dinya kaula nenjo aya hiji sato hanjat ti laut, huluna tujuh tandukna sapuluh. Unggal tanduk make makuta, dina unggal huluna aya tulisan hiji ngaran anu ngahina ka Allah.
mad	(12-18) Naga ganeka laju manjeng e paseser. (13-1) Kaula pas nengale badha keban raja kalowar dhari dhalem tase'. Keban ganeka atandhu' sapolo ban acethak papetto'. E tandhu'na se sapolo ganeka badha jamang settong ebang, ban e saneyap cethagga badha tolesanna, aropa nyama panyeya'an ka Allah.

The third and biggest issue in verse alignment, however, is *verse ranges*. This issue is similar to combined verses but is inconsistent across different languages. The verse ranges issue arises in combined verses, in which the content of the entailing verse only contains the identifier of the current combined verse. Acts 8:28 in the Balinese Bible, for example, is simply "(8-27)", the identifier of the preceding verse, since Acts 8:27 also includes the content of Acts 8:28. However, the Javanese Bible separates these two verses, causing a translation pair discrepancy. Table 5 illustrates this misalignment issue.

Simply removing identifiers will not work in this case because it will still leave the content of the pointing verse to be empty. In order to solve this issue and generate as many valid translation pairs as possible, we address the verse ranges issue by combining the involved verses in all languages, creating a new and final "key" to identify the verse. Table 6 demonstrates this approach.

Therefore, in the previous example, Acts 8:27-28 will be labeled as a single valid translation pair instead of two separate entries. As a result, when aligning two languages, only those with the exact verse key will be considered as a

TABLE 5. Acts 8:28 in Javanese, Indonesian, Sundanese, Madurese, and Balinese translations. Verse identifiers are highlighted in red. Some translations separate two or more verse contents into different verse identifiers, while others combine them with preceding verses – leaving only the trailing verse identifier.

ISO	Verse Content
jav	Nalika samono panjenengane lagi tindak kondur nitih kreta karo maos kitabe Nabi Yesaya.
ind	Sekarang orang itu sedang dalam perjalanan pulang dan duduk dalam keretanya sambil membaca kitab nabi Yesaya.
sun	(8-27)
mad	(8-27)
ban	(8-27)

TABLE 6. Acts 8:28 in Javanese, Indonesian, Sundanese, Madurese, and Balinese translations. Verse identifiers highlighted in red. Some translation separate two or more verse contents into different verse identifiers, while others combine them with preceding verses – leaving only the trailing verse identifier. We combine these verse identifiers to create a new verse key.

ISO	Key	Verse Content
sun	Acts 8:27-28	Pilipus geuwat angkat. Di eta jalan aya hiji pajabat luhur urang Etiopia, anu ngurus harta kakayaan Sri Kandasi ratu nagri Etiopia, tas ti Yerusalem ngalakonan ibadah. Eta pajabat tunggang kareta seja mulih ka nagarana, sajajalan ngaos Kitab Nabi Yesaya.
	~~Acts 8:28~~	~~(8:27)~~
ban	Acts 8:27-28	Irika Dane Pilipus makinkin raris mamargi. Duk punika wenten prakangge agung saking jagat Etiopia nuju mamargi mantuk. Prakangge agung punika, dados prakangge buat ngetangang druen Sri Kandake, Sang Raja Putri ring jagat Etiopia. Dane sampun lunga ka kota Yerusalem ngaturang bakti ring Ida Sang Hyang Widi Wasa, tur sane mangkin dane mawali mantuk nglinggihin kreta. Sajeroning pamargin danene punika, dane ngwacen cakepan dane Nabi Yesaya.
	~~Acts 8:28~~	~~(8:27)~~

TABLE 7. Number of verse-pair samples in each dataset split consisting of Bible verses in various languages of Indonesia.

Config	#train samples	#test samples	#validation samples
Baseline	535,990	154,522	76,490
All	13,168,780	3,871,400	1,894,096

translation pair. Using this strategy, we conducted automatic verse-level alignment as a pre-processing step for our translation data generation.

2) TRANSLATION DATA GENERATION

After performing verse alignment as detailed in the previous sections, we applied the pipeline shown in Fig. 2. We took all of the verse keys found across all of the languages and split them into train (70%), test (20%), and validation (10%) splits. For each split, we then generate all possible permutation pairs of the source and target language. For each pair, we check if the verse key is present in both the source and target corpora, and if so, include it as part of the dataset split. After the permutation of verse pairs, we obtained training splits as shown in Table 7.

TABLE 8. List of languages which are covered in NusaX. Modification of Table 6 found in [12].

ISO 639-3	Language	Dialect
ace	Acehnese	Banda Aceh
ban	Balinese	Lowland
bbc	Toba Batak	Toba, Humbang
bjn	Banjarese	Hulu, Kuala
bug	Buginese	Sidrap
ind	Indonesian	–
jav	Javanese	Matraman
mad	Madurese	Situbondo
min	Minangkabau	Padang, Agam
nij	Ngaju	Kapuas, Kahayan
sun	Sundanese	Priangan

B. SOCIAL MEDIA TEXTS AS LOW-RESOURCE TEXT DOMAIN

Reference [12] provided a high-quality translation dataset consisting of 1,000 translation pairs from 11 languages of Indonesia as outlined in Table 8. From the 1,000 translation pairs, there are 500 training samples, 400 testing samples, and 100 validation samples.

It was originally Indonesian sentiment sentences sourced from social media platforms, which were then translated and annotated by human experts. The dataset, called NusaX, serves as an excellent fine-tuning dataset example in a low-resource setting. Though our initial fine-tuned model may be performant on religious texts, it is very much biased towards the religious domain. Inspired by [24], we will be using social media texts from NusaX, and further specialize our model for translating texts in that domain.

We, therefore, hypothesize that a many-to-many, multilingual translation model trained on an intermediate, mid-resource dataset like religious texts is an effective method to fine-tune to a low-resource, niche domain like social media where data is scarce.

IV. MODEL AND TRAINING SETUP

A. MODEL

Our model was initialized with the pre-trained model checkpoint of mT5 [9], specifically `mt5-base`. The architecture configuration has 12 attention heads, 12 encoder layers, 12 decoder layers, a hidden size of 768, and a feed-forward dimension of 2,048. This is the same model configuration and approach taken by [25] due to the practicality of the computing resource required for training and its adequate performance. We used the HuggingFace [31] implementation of mT5.

Translation is framed as a text-to-text task, with which the model was trained to learn. This follows the proposed method of T5 [7]. Furthermore, like MMTAfrica [25] the task is framed as the following: given a source language X and target language Y, and a corpus of parallel texts $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, we frame the text-to-text task as having an input $<Y_{tag}> x_i$ and a target y_i. Here, Y_{tag} corresponds to the target language ISO 639-3 code.

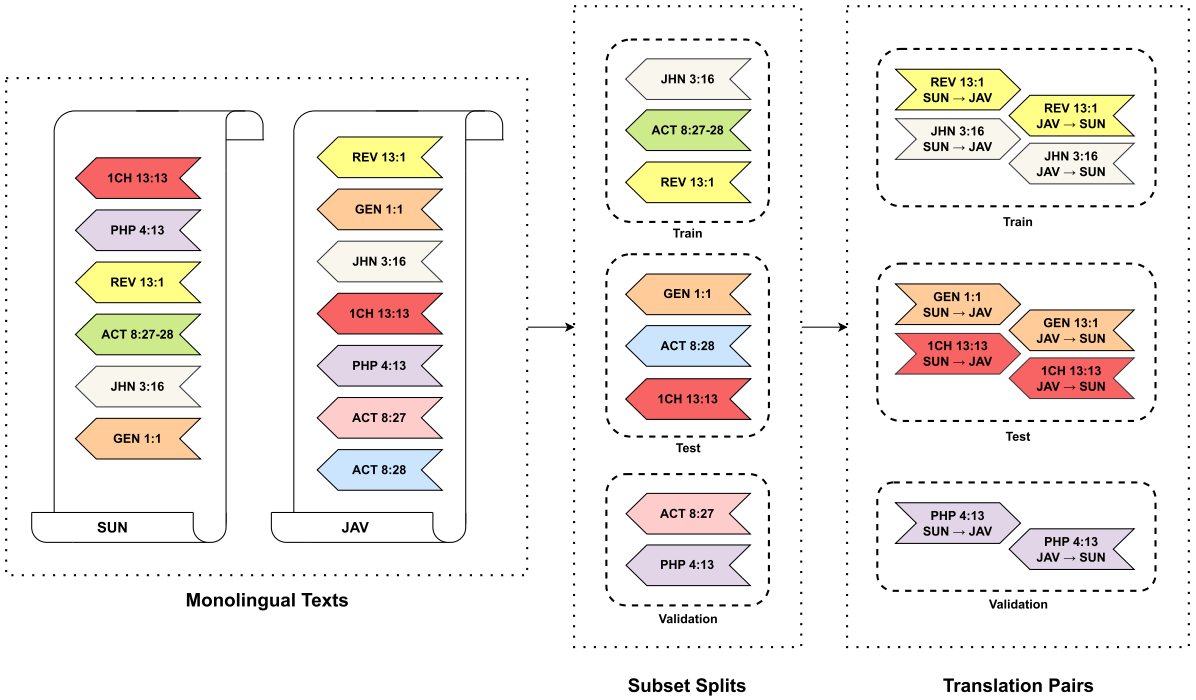


FIGURE 2. Pipeline for religious text translation data generation.

B. TRAINING

There are a few steps to the overall training process. Firstly, we fine-tuned the `mt5-base` checkpoint [9] on religious text translation as an intermediate mid-resource text-domain. In this step, the model learned many-to-many translations using all possible directions of languages found in the training dataset. In `Baseline`, this includes the three languages that mT5 was initially pre-trained on (`ind`, `jav`, `sun`), as well as four new languages of Indonesia (`ace`, `ban`, `bug`, `min`). Afterward, `All` repeats the same process on all 45 languages found in Table 2, most of which are extremely under-resourced.

Then, the newly trained many-to-many multilingual translation model was leveraged to a low-resource setting of social media text translation, using NusaX [12] as training data. Similarly, the `Baseline` model learned many-to-many translation using all languages that it supports (`ace`, `ban`, `bug`, `ind`, `jav`, `min`, `sun`), while `All` used all 11 languages of Indonesia that NusaX covers. From this, we obtained four model checkpoints:

- **Indo-T5**: mT5 fine-tuned on many-to-many translation of religious texts on 7 `Baseline` languages.
- **Indo-T5-NusaX**: Indo-T5 fine-tuned on many-to-many translation of social media texts on 7 `Baseline` languages.
- **Indo-T5-v2**: mT5 fine-tuned on many-to-many translation of religious texts on `All` 45 languages of Indonesia.
- **Indo-T5-v2-NusaX**: Indo-T5-v2 fine-tuned on many-to-many translation of social media texts on 11 languages of Indonesia.

Thus, our overall proposed method is a 2-step fine-tuning, where the first introduces new unseen languages of Indonesia to mT5 and the second further specializes the model's capability to translate texts in a niche, low-resource text-domain.

Furthermore, to compare the effectiveness of multilingual fine-tuning versus bilingual fine-tuning, we also fine-tuned both Indo-T5 and Indo-T5-v2 on bilingual pairs of languages of NusaX. Hence, instead of having one model that does multilingual, many-to-many translation, we trained several translation models which are strictly bilingual and unidirectional.

C. EVALUATION

In order to evaluate the effectiveness of our proposed method and the overall performance of our translation model, we will be benchmarking our method's results on the NusaX test set and compare them to benchmark models presented in [12], as well as zero-shot results of NLLB [6], [29], [30]. To quantify our model's performance on translation, we will use the SacreBLEU metric [32], which is an improvement of the original BLEU metric [33] and a beam size of 10 during decoding. All evaluation will be done in the direction of `ind` → `x` and `x` → `ind` where `x` is one of the regional languages of Indonesia.

Moreover, the evaluation will be categorized into two: `Baseline` languages and `All` languages. We do this on purpose to investigate the effects of the *curse of multilinguality* as observed in cross-lingual models like XLM-RoBERTa [34]. Intuitively, fitting all 45 languages into mT5 as opposed to

TABLE 9. Hyperparameters used for fine-tuning.

Model	Dataset	Batch Size	#optim	LR	λ
Indo-T5	Alkitab SABDA	4,096	5,000	$8e^{-4}$	0
Indo-T5-NusaX	NusaX	256	500	$2e^{-4}$	0
Indo-T5-NusaX	NusaX (Bilingual)	32	500	$1e^{-3}$	0.01
Indo-T5-v2	Alkitab SABDA	4,096	10,000	$8e^{-4}$	0
Indo-T5-v2-NusaX	NusaX	256	1,000	$2e^{-4}$	0
Indo-T5-v2-NusaX	NusaX (Bilingual)	32	500	$1e^{-3}$	0.01

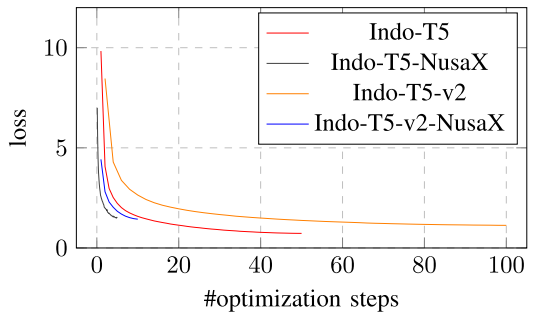


FIGURE 3. Training loss graph of our fine-tuned models.

only 7 initial languages will lead to a deterioration in performance. We investigated whether this phenomenon occurs through the following experiments.

V. EXPERIMENTS

As mentioned in §IV-A, we used the HuggingFace [31] implementation of `mt5-base` [9] in all of our experiments. Furthermore, we used the PyTorch [35] implementation of the fused AdamW optimizer ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e^{-8}$) [36], alongside a learning rate scheduler that decays linearly after several warm-up steps. The details of our optimization hyperparameters are listed in Table 9.

All of our experiments were conducted on 8 NVIDIA A100 40GB GPUs and utilized bfloat16 precision during training. A plot of our training loss graph is shown in Fig. 3.

A. BASELINE LANGUAGES

Like [25], we wanted to investigate whether languages found in the pre-training corpus of mT5 have an impact on the downstream translation performance of the model. Hence, we started with a smaller subset of languages which we term as `Baseline`.

Languages included in the pre-training corpus of mT5 are `ind`, `jav`, and `sun`, while `ace`, `ban`, `bug`, and `min`, were previously unseen by the model. We chose this specific subset of languages since they are all supported by IndoBART and IndoGPT [11], NLLB [6], [29], [30], and NusaX [12].

We first fine-tuned `mt5-base` as a multilingual, many-to-many translation model called Indo-T5 on religious texts from various translations of the Bible. We used a batch size of 256 sentences, with 16 gradient accumulation steps, totaling an effective batch size of 4,096. The model was trained for 5,000 optimization steps, which is equivalent to about 38 epochs, with a learning rate of $8e^{-4}$ with 500 warm-up steps.

TABLE 10. SacreBLEU scores of Indo-T5 on translation from Indonesian to regional languages (`ind` → `x`) on religious texts from Alkitab SABDA.

	ind → x						
Model	ace	ban	bug	jav	min	sun	**avg**
Indo-T5	18.59	20.62	17.16	27.31	21.74	12.93	19.73

Afterward, we further fine-tuned Indo-T5 on NusaX [12] which contains texts sourced from various social media platforms. Like the previous experiment, the model was fine-tuned as a multilingual translation model called Indo-T5-NusaX. Specifically, we used a batch size of 256 sentences and trained for 500 optimization steps with a learning rate of $2e^{-4}$ with 50 warm-up steps.

In addition, we also fine-tuned Indo-T5 as bilingual translation models on NusaX to compare the performance of Indo-T5-NusaX against its bilingual counterparts. Because there are 7 languages in `Baseline`, we end up with 12 bilingual, unidirectional models. They were all fine-tuned with a batch size of 32 sentences and were similarly trained for 500 optimization steps with a learning rate of $1e^{-3}$ with 50 warm-up steps.

B. ALL LANGUAGES

We then repeated the same processes as outlined in §V-A with all 45 languages shown in Table 2. Thus, we introduced 43 new languages which were previously unseen to mT5 through fine-tuning on religious texts, i.e. `All` languages.

Like in `Baseline`, we fine-tuned `mt5-base` as a multilingual, many-to-many translation model called Indo-T5-v2 on religious texts. This followed the same setup as Indo-T5, with the exception of training for 10,000 optimization steps with 1,000 warm-up steps.

Moreover, following the exact setup as Indo-T5-NusaX, we similarly took Indo-T5-v2 and fine-tuned it on all 11 languages of Indonesia as covered in NusaX [12] shown in Table 8, with the exception of training for 1,000 optimization steps with 1,000 warm-up steps. The resultant model is called Indo-T5-v2-NusaX.

Finally, we fine-tuned Indo-T5-v2 as bilingual translation models on NusaX for comparison purposes. We thus obtained 20 bilingual, unidirectional models from the 11 languages of Indonesia found in NusaX. They also followed an identical setup to that of the `Baseline` bilingual models.

VI. RESULTS AND ANALYSIS

We conducted an evaluation of our models as detailed in §IV-C, where we compared our models' results with those shown in [12] and the zero-shot capability of NLLB [6], [29], [30].

A. BASELINE LANGUAGES RESULTS

The results of Indo-T5 on religious texts on `Baseline` languages are listed in Table 10 and Table 11.

TABLE 11. SacreBLEU scores of Indo-T5 on translation from regional languages to Indonesian (x → ind) on religious texts from Alkitab SABDA.

	x → ind						
Model	ace	ban	bug	jav	min	sun	**avg**
Indo-T5	21.74	19.37	19.66	34.99	22.41	17.3	22.58

We observed that our multilingual model is generally more capable to translate from regional languages to Indonesian as shown with the higher SacreBLEU score in the direction of x → ind. Furthermore, our model performed best on jav in either direction. We hypothesize that this is because both Indonesian and Javanese were included in the pre-training corpus of mT5 [9]. However, we also noticed that the model performed worst on Sundanese despite the presence of the language in mT5's pre-training corpus. Similarly, the model showed decent results on translations to and from Minangkabau, even though the language was previously unseen to the model. High lexical overlap between Minangkabau and Indonesian as pointed out by [37] may be a strong reason for this discrepancy.

Table 12 and Table 13 contain the results of Indo-T5-NusaX which was fine-tuned on Baseline languages of NusaX [12]. We marked languages that were excluded from Baseline with a dash (-) and compared the average results on the remaining languages only. We found that bilingual fine-tuning resulted in the best-performing models, achieving on-par if not higher SacreBLEU scores than the benchmark models. Most notably, our proposed method outperformed mT5 Base, which is a strong indication that fine-tuning only on social media data leads to a worse result, mainly due to the small size of the translation dataset. We also noticed, as we did in religious texts fine-tuning, that our models are more performant in the direction of x → ind. Multilingual fine-tuning, however, only showed similar capabilities to that of generic multilingual models like mBART [10] and mT5 [9], only occasionally outperforming them.

B. ALL LANGUAGES RESULTS

We presented the results of Indo-T5-v2 on religious texts on All languages on Table 16. We remark that our model only performs slightly better in the direction of x → ind, unlike previous results where the difference was more noticeable. Likewise, while our model still performed strongly in Javanese, it showed decent results in previously unseen languages such as Iban (iba), Batak Simalungun (bts), and Manggarai (mqy) where SacreBLEU > 25.0 in particular directions.

Additionally, we also observed that the number of verses – and correspondingly the number of training pairs – do not affect the translation capability of the model. Our model displayed on-par performances in languages that had both the Old and New Testament and those which only had New Testament. We suspect that language similarities instead have a stronger effect on the performance of our models in certain languages, which we will investigate in the subsequent sections.

Results of Indo-T5-v2-NusaX on All languages of NusaX [12] are shown in Table 12 and Table 13. We noticed the same trend of results as we did in Baseline languages. Namely, bilingual fine-tuning resulted in the highest overall SacreBLEU scores, even on languages that were previously unseen to the model. For instance, our model has the highest score on ind → mad, mad → ind, and bbc → ind, outperforming benchmark models.

We again observe that our model is generally more capable of translating in the direction of x → ind, even if this was not the case in religious texts. In the same way, multilingual fine-tuning on NusaX resulted in a lackluster performance compared to its counterpart bilingual benchmark models. Results when compared to Indo-T5-NusaX also confirmed the curse of multilinguality, where fitting more languages into one model of the same size will result in a decrease in overall performance. The only case where our model showed a better performance when fine-tuning on all 45 languages is on jav → ind.

C. PRE-TRAINED LANGUAGES AND TRANSFER LEARNING

We observed that our model is consistently performant on Javanese, which suggests that its inclusion in the pre-training corpus of mT5 [9] was crucial to its downstream translation performance. However, we do not observe a similar performance on Sundanese, even though the language was also included in the pre-training corpus, mC4 [7]. We attribute the poorer performance of our fine-tuned model on Sundanese due to the significantly smaller amount of pre-training corpus available for Sundanese (280,719 sentences) compared to Javanese (581,528 sentences) in mC4 [38]. Nevertheless, [38] also noted that the quality of both Javanese and Sundanese corpora found in mC4 is extremely poor.

Furthermore, almost all of our experiments showed that translation in the direction of x → ind resulted in a higher SacreBLEU score than in the opposite direction. We hypothesize that this is due to the decoder's ability and familiarity with Indonesian, as it is one of the largest languages in mC4 [7]. On the contrary, most of the regional languages are new to mT5. For previously unseen languages, [25] suggested further fine-tuning mT5 on these languages with a masked language modeling-like task to help the model learn a better representation of these languages and thereby improve the downstream translation capability of the model.

Moreover, we saw a significant improvement in SacreBLEU scores when doing a two-step fine-tuning on translation, leading to an increase of about 8-10 SacreBLEU scores on NusaX [12], when compared with a one-step fine-tuning of mt5-base. This strongly indicates that firstly fine-tuning on a widely available, mid-resource text domain like Bible translations aids the performance of the model on a low-resource translation domain like social media texts. This is in line with [24] who similarly found that religious texts serve as

TABLE 12. SacreBLEU scores on translation from Indonesian to regional languages (ind → x). Results of benchmark models were obtained from Table 4 in [12].

		ind → x										
Model	**#params**	ace	ban	bbc	bjn	bug	jav	mad	min	nij	sun	**avg**
IndoGPT	117M	9.60	14.17	8.20	22.23	5.18	24.05	14.44	26.95	17.56	23.15	16.55
IndoBART v2	132M	**19.21**	**27.08**	**18.41**	**40.03**	**11.06**	**39.97**	28.95	48.48	**27.11**	**38.46**	29.88
mBART-50 Large	610M	17.21	22.67	17.79	34.26	10.78	35.33	28.63	43.87	25.91	31.21	26.77
mT5 Base	580M	14.79	18.07	18.22	38.64	6.68	33.48	0.96	45.84	13.59	33.79	22.41
NLLB-200 Distilled	600M	2.74	4.87	-	-	1.66	17.66	-	9.79	-	11.92	8.11
Indo-T5 NusaX Multilingual	580M	16.02	22.48	-	-	8.86	33.65	-	33.65	-	29.76	24.07
Indo-T5 NusaX Bilingual	580M	17.99	27.03	-	-	10.80	39.63	-	**51.56**	-	35.16	**30.36**
Indo-T5 v2 NusaX Multilingual	580M	14.28	19.19	14.86	28.39	8.05	28.70	20.95	32.70	22.30	26.19	21.56
Indo-T5 v2 NusaX Bilingual	580M	17.58	24.24	16.69	38.81	10.20	37.87	**29.77**	50.90	26.93	34.22	28.72

TABLE 13. SacreBLEU scores on translation from regional languages to Indonesian (x → ind). Results of benchmark models were obtained from Table 5 in [12].

		x → ind										
Model	**#params**	ace	ban	bbc	bjn	bug	jav	mad	min	nij	sun	**avg**
IndoGPT	117M	7.01	13.23	5.27	19.53	1.98	27.31	13.75	23.03	10.83	23.18	14.51
IndoBART v2	132M	24.44	40.49	19.94	**47.81**	12.64	**50.64**	36.10	58.38	**33.50**	**45.96**	36.99
mBART-50 Large	610M	18.45	34.23	17.43	41.73	10.87	39.66	32.11	59.66	29.84	35.19	31.92
mT5 Base	580M	18.59	21.73	12.85	42.29	2.64	45.22	32.35	58.65	25.61	36.58	29.65
NLLB-200 Distilled	600M	8.11	21.24	-	-	6.18	30.54	-	40.49	-	26.91	22.46
Indo-T5 NusaX Multilingual	580M	23.94	35.30	-	-	**16.68**	29.76	-	48.10	-	36.54	31.72
Indo-T5 NusaX Bilingual	580M	**24.78**	**42.15**	-	-	16.27	47.26	-	**62.94**	-	42.39	**39.30**
Indo-T5 v2 NusaX Multilingual	580M	21.01	30.43	18.57	34.21	14.42	35.19	27.04	42.64	26.90	33.78	28.42
Indo-T5 v2 NusaX Bilingual	580M	22.87	39.48	**20.48**	44.53	15.97	45.20	**36.65**	60.97	32.38	39.80	35.83

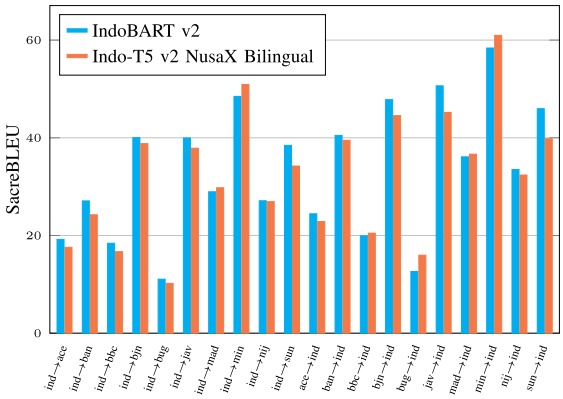


FIGURE 4. SacreBLEU scores of IndoBART v2 [11] and Indo-T5 v2 NusaX Bilingual on NusaX [12].

an effective intermediate text domain for translation before specializing the model on a more niche text domain.

We also noticed a similar trend in SacreBLEU scores between our model and that of IndoBART [11], as shown in Fig. 4. For instance, both models were performant on Balinese, Banjarese, Madurese, and Minangkabau even though they were all not found in either of the two models' pre-training corpora. Reference [12] proposed that similarities between language lexicons and the close relation among languages in the Malayo-Polynesian language family led to a positive transfer in translation performance.

As an example, Banjarese is similar to Malay [39] and has a 73% lexical similarity with Indonesian [12], both languages included in mC4. Minangkabau and Indonesian also have commonalities in their vocabulary and syntax [37]. In contrast, on languages like Buginese and Toba Batak, both IndoBART and our model exhibit poor performances, confirming the empirical findings of [12] that attributed the low vocabulary overlap between either language with Indonesian [1] to be the reason for their underwhelming results.

D. ZERO-SHOT TRANSLATION

We examined the zero-shot capability of our multilingual model, namely Indo-T5-v2-NusaX, on the direction of eng → x. English texts make up the largest percentage of the pre-training corpus of mT5 [9]. We are interested to find out whether our model which has been fine-tuned exclusively on regional languages of Indonesia would be able to transfer its capability to translate English texts. We calculated the SacreBLEU scores on the test subset of NusaX [12]. Table 14 shows the zero-shot result of our model in the direction of eng → ind, compared to supervised benchmark models as conducted by [12]. Table 15 similarly lists zero-shot results of our model in the direction of eng → x, where x is a regional language of Indonesia.

TABLE 14. SacreBLEU scores of Indo-T5-v2-NusaX on zero-shot translation from English to Indonesian (eng → ind).

Model	eng → ind
Supervised	
IndoGPT	4.26
IndoBART v2	11.73
mBART-50 Large	17.92
mT5 Base	12.96
Zero-shot	
Indo-T5 v2 NusaX	5.51

TABLE 15. SacreBLEU scores of Indo-T5-v2-NusaX on zero-shot translation from English to regional languages of Indonesia (eng → x).

eng → x									
ace	ban	bbc	bjn	bug	jav	mad	min	nij	sun
2.93	2.63	3.46	3.96	2.58	4.60	3.04	4.17	3.39	4.22

We observed that the SacreBLEU scores of zero-shot translation are significantly lower than their supervised counterparts. Nonetheless, in the direction of eng → ind, our model was able to outperform IndoGPT, which was trained in a supervised manner.

In addition, we also noticed that our best zero-shot target languages were Javanese (jav), Sundanese (sun), Minangkabau (min), and Banjarese (bjn), on which our model also achieved the highest SacreBLEU scores when translating from Indonesian, as shown in Table 12.

We suspect that pre-trained languages of mT5 and lexical similarity to Indonesian also highly impacted these zero-shot results, in line with our analysis in §VI-C.

We also present several cherry-picked results of zero-shot eng → ind translations in Table 17, which indicates that our model tends to generate code-mixed translations while also potentially modifying the text's actual meaning.

E. TRANSLATION BIAS SHIFT

We explored shifts in resultant translation text bias before and after fine-tuning on social media texts, as similarly conducted by [24]. Table 18 shows two translated Bible verses as inferred by Indo-T5-v2 trained on religious texts and Indo-T5-v2-NusaX further fine-tuned on social media texts.

Both models were able to capture the meaning of the intended text, but Indo-T5-v2 used religious terms such as *firman* (the word of God) and *kitab* (holy book) while Indo-T5-v2-NusaX used more neutral terms such as *kabar* (message, news, report) and *buku* (book), respectively.

VII. CONCLUSION

Based on our experiments, we conclude that the many-to-many translation model developed for the languages of Indonesia using mT5 as a baseline checkpoint is a promising approach for low-resource language pairs. The two-step, intermediate translation domain approach utilizing religious texts from the Bible proved to be effective in achieving better performance in low-resource language pairs.

Likewise, we also found that further fine-tuning the multilingual intermediate translation model as a bilingual translation model on the low-resource domain improves its overall performance as opposed to fine-tuning it as a multilingual translation model again. On particular language pairs, this approach led to an increase in SacreBLEU scores ranging from +0.84 (ace → ind) to +18.33 (min → ind).

TABLE 16. SacreBLEU scores of Indo-T5-v2 on translation from Indonesian to regional languages (ind → x) and regional languages to Indonesian (x → ind) on religious texts from Alkitab SABDA.

ISO	Language	ind → x	x → ind
	Old Testament and New Testament		
ban	Balinese	15.68	14.48
bbc	Toba Batak	22.18	20.79
bts	Batak Simalungun	25.03	22.97
btx	Batak Karo	12.64	12.99
bug	Buginese	13.75	14.61
jav	Javanese	20.72	28.40
mad	Madurese	13.58	13.83
mak	Makassarese	13.07	13.85
sda	Toraja-Sa'dan	19.17	21.23
sun	Sundanese	8.65	12.11
	New Testament only		
abs	Ambon	18.13	21.22
ace	Aceh	16.94	18.97
atq	Aralle Tabulahan	15.26	16.57
bkl	Berik	13.56	13.94
blz	Balantak	21.10	19.10
btd	Pakpak Dairi	15.24	18.07
bvz	Bauzi	10.71	10.64
gbi	Galela	14.66	14.23
gor	Gorontalo	21.36	20.30
hvn	Sabu	22.12	18.33
iba	Iban	28.20	21.42
kgr	Abun	18.27	13.61
kzf	Kaili Daa	18.81	17.78
ljp	Lampung	20.04	19.95
mej	Meyah	17.53	12.92
min	Minangkabau	19.66	19.17
mkn	Kupang	16.02	14.41
mog	Mongondow	12.52	18.71
mqj	Mamasa	16.11	16.98
mqy	Manggarai	23.39	25.20
mvp	Duri	17.34	16.57
mwv	Mentawai	15.42	16.15
nia	Nias	15.61	17.60
nij	Ngaju	19.66	19.15
npy	Napu	16.44	16.11
pmf	Taa	17.48	15.79
ppk	Uma	15.02	16.56
ptu	Bambam	15.76	16.67
sas	Sasak	18.15	21.53
sxn	Sangir	12.32	18.31
tby	Tabaru	18.48	16.65
twu	Rote	19.13	17.35
yli	Yali Angguruk	9.42	11.89
yva	Yawa	13.86	13.97
	avg	17.00	17.30

TABLE 17. Cherry-picked zero-shot `eng` → `ind` translation results of Indo-T5-v2-NusaX on NusaX texts.

Source	Target	Prediction
Near the hotel I stayed in, reachable by foor, so many food choice here, the place is huge, and fun	Dekat dengan hotel saya menginap, hanya ditempuh jalan kaki, di sini banyak sekali pilihan makanannya, tempat yang luas, dan menyenangkan	Menyeberang hotel I duduk di sana, menarik pesan, bahkan banyak pilihan makanan di sini, tempatnya luas, sangat nyaman
Located inside the Braga City Walk, which is in the same building as Aston and Fave Hotel, this is the perfect hangout spot. The coffee-tea mix that I tried for the first time was actually amazing. Combined with a sunny-side up egg and you got yourself the perfect meal for chatting with your friends. The smoke-free zone just adds to the comforting feeling as you watch the view of people coming and going in this mall	Bertempat di braga city walk yang satu gedung dengan aston dan fave hotel, tempat ini sangat nyaman buat kongkow-kongkow. Kopi campur teh yang baru pertama kali saya nikmati ternyata sangat enak, dipadu dengan telur setengah matang menjadi pendamping mengobrol bersama teman-teman. Area yang bebas merokok semakin mengasyikkan sambil menikmati pemandangan lalu lalang orang-orang yang keluar masuk mal ini.	Lokasi di belakang bandung citywalk, yang terletak di sebelah bangunan aston dan fav hotel, ini adalah yang paling romantis. Produksi kopi yang dipilih dari permulaan waktu memang sangat indah. Dibandingkan dengan sunny valley egg dan engkau mendapatkan sendiri yang paling terbaik makanan untuk nongkrong bareng teman-teman. Smoke free food hanya menambah kualitas hati karena engkau melihat pemandangan orang pergi dan jalan di mall ini.
YouTube is showing more ads the more days pass, it's annoying!	Youtube semakin hari semakin banyak iklannya, menyebalkan!	YouTube sudah membuat banyak adset di berapa hari lamanya, sudah mengecewakan!

TABLE 18. Comparison of translation of religious texts by models trained on religious texts and casual texts respectively.

`jav→ind`	
Source	Ana ing kono karone padha nggelarake pangandika ana ing kutha Perga, banjur tindak menyang ing kutha Atalia, ing pesisir.
Target	Di situ mereka memberitakan firman di Perga, lalu pergi ke Atalia, di pantai.
Indo-T5 v2	Di sana mereka memberitahukan firman di Perga, lalu pergi ke Atalia, di pantai.
Indo-T5 v2 NusaX	Di sana mereka memberitahukan kabar di Perga, lalu pergi ke Atalia, di pinggir selera.
`sun→ind`	
Source	Hayang teh ieu omongan kuring aya nu nyatetkeun, tiap kecapna ditulis dijieun buku.
Target	Ah, kiranya perkataanku ditulis, dicatat dalam kitab,
Indo-T5 v2	Maksudku ialah, supaya aku menuliskan perkataan-perkataan ini dan menuliskannya dalam suatu kitab.
Indo-T5 v2 NusaX	Maulah ini ulasan saya ada yang menentuknya, setiap kecapna ditulis di dalam buku

TABLE 19. English translations of Indonesian texts found in Fig. 1, Table 17, and Table 18.

Indonesian	English
Pada mulanya Allah menciptakan langit dan bumi. (Genesis 1:1 TB)	In the beginning God created the heavens and the earth. (Genesis 1:1 NIV)
Pada mulanya, waktu Allah mulai menciptakan alam semesta. (Genesis 1:1 BIS)	In the beginning, when God created the universe.
Mau bikin postingan yang isinya mengedukasi customers gojek.	Want to make a post that educates gojek customers.
Ingin buat posting yang isinya mengajar pelanggan gojek.	Want to make a post that teaches gojek customers.
Menyeberang hotel I duduk di sana, menarik pesan, bahkan banyak pilihan makanan di sini, tempatnya luas, sangat nyaman	Across from hotel I was sitting there, interesting message, even a lot of food choices here, the place is wide, very comfortable
Lokasi di belakang bandung citywalk, yang terletak di sebelah bangunan aston dan fav hotel, ini adalah yang paling romantis. Produksi kopi yang dipilih dari permulaan waktu memang sangat indah. Dibandingkan dengan sunny valley egg dan engkau mendapatkan sendiri yang paling terbaik makanan untuk nongkrong bareng teman-teman. Smoke free food hanya menambah kualitas hati karena engkau melihat pemandangan orang pergi dan jalan di mall ini.	The location is behind bandung citywalk, which is next to the aston building and fav hotel, is the most romantic. The coffee production selected from the beginning of time is indeed very beautiful. Compare with sunny valley egg and you get yourself the most best food for hanging out with friends. Smoke free food only adds to the quality of the heart because you see the sight of people going and walking in this mall.
YouTube sudah membuat banyak adset di berapa hari lamanya, sudah mengecewakan!	YouTube already made many adset in how many days ago, already disappointing!
Di situ mereka memberitakan firman di Perga, lalu pergi ke Atalia, di pantai. (Acts 14:25 TB)	and when they had preached the word in Perga, they went down to Attalia. (Acts 14:25 NIV)
Di sana mereka memberitahukan firman di Perga, lalu pergi ke Atalia, di pantai.	There they gave news in Perga, then went to Atalia, on the beach.
Di sana mereka memberitahukan kabar di Perga, lalu pergi ke Atalia, di pinggir selera.	There they gave news to Perga, then went to Atalia, on the edge of the city.
Ah, kiranya perkataanku ditulis, dicatat dalam kitab, (Job 19:23 TB)	"Oh, that my words were recorded, that they were written on a scroll, (Job 19:23 NIV)
Maksudku ialah, supaya aku menuliskan perkataan-perkataan ini dan menuliskannya dalam suatu kitab.	I mean, so that I write down these words and write them in a book.
Maulah ini ulasan saya ada yang menentuknya, setiap kecapna ditulis di dalam buku	But this is my review that determines it, every word is written in the book

However, since we introduced significantly more languages while keeping the same number of parameters, we also observed the phenomenon curse of multilinguality which led to lower SacreBLEU scores in several languages as seen in Table 12 and Table 13.

Our findings suggest that leveraging existing multilingual sequence-to-sequence language models for low-resource languages is a viable approach, and can possibly lead to improvements in machine translation quality for under-resourced languages. Multilingual translation fine-tuning from a multilingual language model also showed a potential for zero-shot translation. Further research is needed to explore the feasibility of this approach for other low-resource languages, as well

TABLE 20. English New International Version (NIV) verse translations of verses found in Table 3, Table 4, Table 5, and Table 6.

Verse Number	Verse Content
Exodus 6:1	Then the Lord said to Moses, "Now you will see what I will do to Pharaoh: Because of my mighty hand he will let them go; because of my mighty hand he will drive them out of his country."
Revelations 13:1	The dragon stood on the shore of the sea. And I saw a beast coming out of the sea. It had ten horns and seven heads, with ten crowns on its horns, and on each head a blasphemous name.
Acts 8:27	So he started out, and on his way he met an Ethiopian eunuch, an important official in charge of all the treasury of the Kandake (which means "queen of the Ethiopians"). This man had gone to Jerusalem to worship,
Acts 8:28	and on his way home was sitting in his chariot reading the Book of Isaiah the prophet.

as to investigate the potential of using alternative intermediate mid-resource translation domains.

APPENDIX ENGLISH TRANSLATIONS

We provide English translations of non-English texts found in our manuscript in Table 19 and Table 20.

ACKNOWLEDGMENT

The authors would like to thank Bina Nusantara University for facilitating and supporting this entire research process, and also would like to thank Bookbot Pty Ltd. for providing access to their Google Cloud Platform Compute Engine resources.

REFERENCES

[1] A. F. Aji, G. I. Winata, F. Koto, S. Cahyawijaya, A. Romadhony, R. Mahendra, K. Kurniawan, D. Moeljadi, R. E. Prasojo, T. Baldwin, J. H. Lau, and S. Ruder, "One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 7226–7249. [Online]. Available: https://aclanthology.org/2022.acl-long.500

[2] A. I. Fauzi and D. Puspitorini, "Dialect and identity: A case study of Javanese use in WhatsApp and line," *IOP Conf. Ser., Earth Environ. Sci.*, vol. 175, Jul. 2018, Art. no. 012111.

[3] Bahasa Kita. (Mar. 2019). *Depdiknas Terbitkan Peta Bahasa*. [Online]. Available: https://www.bahasakita.com/id/bahas-bahasa/depdiknas-terbitkan-peta-bahasa/

[4] Badan Pengembangan Bahasa dan Perbukuan. (2019). *Bahasa Daerah Di Indonesia*. [Online]. Available: https://dapobas.kemdikbud.go.id/homecat.php?show=url/petabahasa

[5] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, and A. Joulin, "Beyond English-centric multilingual machine translation," *J. Mach. Learn. Res.*, vol. 22, no. 1, pp. 4839–4886, 2021.

[6] N. Team et al., "No language left behind: Scaling human-centered machine translation," 2022, *arXiv:2207.04672*.

[7] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 5485–5551, 2020.

[8] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 7871–7880. [Online]. Available: https://aclanthology.org/2020.acl-main.703

[9] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "MT5: A massively multilingual pre-trained text-to-text transformer," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.* Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 483–498. [Online]. Available: https://aclanthology.org/2021.naacl-main.41

[10] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 726–742, Dec. 2020. [Online]. Available: https://aclanthology.org/2020.tacl-1.47

[11] S. Cahyawijaya, G. I. Winata, B. Wilie, K. Vincentio, X. Li, A. Kuncoro, S. Ruder, Z. Y. Lim, S. Bahar, M. Khodra, A. Purwarianti, and P. Fung, "IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 8875–8898. [Online]. Available: https://aclanthology.org/2021.emnlp-main.699

[12] G. I. Winata, A. F. Aji, S. Cahyawijaya, R. Mahendra, F. Koto, A. Romadhony, K. Kurniawan, D. Moeljadi, R. E. Prasojo, P. Fung, T. Baldwin, J. H. Lau, R. Sennrich, and S. Ruder, "NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages," in *Proc. 17th Conf. Eur. Chapter Assoc. Comput. Linguistics*. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 815–834. [Online]. Available: https://aclanthology.org/2023.eacl-main.57

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[14] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, May 2015, pp. 1–15.

[15] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2023, *arXiv:1606.08415*.

[16] Y. Tay, D. Bahri, D. Metzler, D.-C. Juan, Z. Zhao, and C. Zheng, "Synthesizer: Rethinking self-attention for transformer models," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10183–10192.

[17] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *J. Mach. Learn. Res.*, vol. 23, no. 1, pp. 5232–5270, 2022.

[18] N. Ng, K. Yee, A. Baevski, M. Ott, M. Auli, and S. Edunov, "Facebook FAIR's WMT19 news translation task submission," in *Proc. 4th Conf. Mach. Transl.* Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 314–319. [Online]. Available: https://aclanthology.org/W19-5333

[19] J. Tiedemann and S. Thottingal, "OPUS-MT—Building open translation services for the world," in *Proc. 22nd Annu. Conf. Eur. Assoc. Mach. Transl. (EAMT)*, Lisbon, Portugal, 2020, pp. 479–480.

[20] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. F. Aji, N. Bogoychev, A. F. T. Martins, and A. Birch, "Marian: Fast neural machine translation in C++," in *Proc. ACL, Syst. Demonstrations*. Melbourne, VIC, Australia: Association for Computational Linguistics, Jul. 2018, pp. 116–121. [Online]. Available: http://www.aclweb.org/anthology/P18-4020

[21] R. Aharoni, M. Johnson, and O. Firat, "Massively multilingual neural machine translation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.* Minneapolis, MN, USA: Association for Computational Linguistics, Jun. 2019, pp. 3874–3884. [Online]. Available: https://aclanthology.org/N19-1388

[22] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI, CA, USA, 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.* Minneapolis, MN, USA: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[24] D. Adelani et al., "A few thousand translations go a long way! Leveraging pre-trained models for African news translation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.* Seattle, WA, USA: Association for Computational Linguistics, Jul. 2022, pp. 3053–3070. [Online]. Available: https://aclanthology.org/2022.naacl-main.223

[25] C. C. Emezue and B. F. P. Dossou, "MMTAfrica: Multilingual machine translation for African languages," in *Proc. 6th Conf. Mach. Transl.* Stroudsburg, PA, USA: Association for Computational Linguistics, Nov. 2021, pp. 398–411. [Online]. Available: https://aclanthology.org/2021.wmt-1.48

[26] A. Ghafoor, A. S. Imran, S. M. Daudpota, Z. Kastrati, Abdullah, R. Batra, and M. A. Wani, "The impact of translating resource-rich datasets to low-resource languages through multi-lingual text processing," *IEEE Access*, vol. 9, pp. 124478–124490, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9529190

[27] K. Vincentio and D. Suhartono, "Automatic question generation using RNN-based and pre-trained transformer-based models in low resource Indonesian language," *Informatica*, vol. 46, no. 7, pp. 103–118, 2022.

[28] D. M. Eberhard, G. F. Simons, and C. D. Fennig, "Ethnologue: Languages of the world," Ethnologue, Dallas, TX, USA, Tech. Rep., 2022. [Online]. Available: https://www.ethnologue.com/

[29] N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, and A. Fan, "The Flores-101 evaluation benchmark for low-resource and multilingual machine translation," *Trans. Assoc. Comput. Linguistics*, vol. 10, pp. 522–538, May 2022. [Online]. Available: https://aclanthology.org/2022.tacl-1.30

[30] F. Guzmán, P.-J. Chen, M. Ott, J. Pino, G. Lample, P. Koehn, V. Chaudhary, and M. Ranzato, "The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*. Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 6098–6111. [Online]. Available: https://aclanthology.org/D19-1632

[31] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*. Stroudsburg, PA, USA: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.6

[32] M. Post, "A call for clarity in reporting BLEU scores," in *Proc. 3rd Conf. Mach. Transl., Res. Papers*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 186–191. [Online]. Available: https://aclanthology.org/W18-6319

[33] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2001, pp. 311–318.

[34] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, Jul. 2020, pp. 8440–8451. [Online]. Available: https://aclanthology.org/2020.acl-main.747

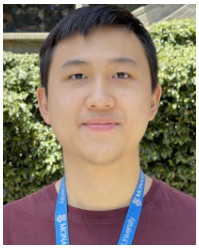
[35] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2019, pp. 1–12.

[36] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. 7th Int. Conf. Learn. Represent. (ICLR)*, New Orleans, LA, USA, May 2019, pp. 1–19. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

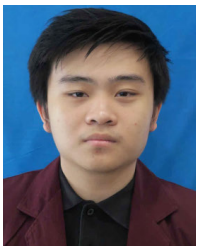
[37] F. Koto and I. Koto, "Towards computational linguistics in Minangkabau language: Studies on sentiment analysis and machine translation," in *Proc. 34th Pacific Asia Conf. Lang., Inf. Comput.* Hanoi, Vietnam: Association for Computational Linguistics, Oct. 2020, pp. 138–148. [Online]. Available: https://aclanthology.org/2020.paclic-1.17

[38] J. Kreutzer et al., "Quality at a glance: An audit of web-crawled multilingual datasets," *Trans. Assoc. Comput. Linguistics*, vol. 10, pp. 50–72, Jan. 2022. [Online]. Available: https://aclanthology.org/2022.tacl-1.4

[39] A. H. Nasution, Y. Murakami, and T. Ishida, "Plan optimization to bilingual dictionary induction for low-resource language families," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 20, no. 2, pp. 1–28, Mar. 2021, doi: 10.1145/3448215.



WILSON WONGSO is currently pursuing the bachelor's degree in computer science with Bina Nusantara University, Indonesia. He is also a machine learning engineer, specializing in natural and speech language processing. His research interests include natural language processing for low-resource languages and Indonesia-related languages.



ANANTO JOYOADIKUSUMO is currently pursuing the bachelor's degree in computer science with Bina Nusantara University, Indonesia. He is also a machine learning engineer and also works as a part-time database administrator. His research interests include NLP, vision, and the deployment-focused text-to-speech.

BRANDON SCOTT BUANA is currently pursuing the bachelor's degree in computer science with Bina Nusantara University, Indonesia. He most recently interned as an AI engineer, with a focus on computer vision and model deployment. His research interests include natural language processing and computer vision.

DERWIN SUHARTONO (Member, IEEE) received the Ph.D. degree in computer science from Universitas Indonesia, in 2018. He is currently a Faculty Member of Bina Nusantara University, Indonesia. His research interest includes natural language processing. Recently, he is continually doing research in argumentation mining and personality recognition. He actively involves in Indonesia Association of Computational Linguistics (INACL), a National Scientific Association in Indonesia, IndoCEISS, and Aptikom. He has his professional memberships in ACM, INSTICC, and IACT. He also takes role as a reviewer in several international conferences and journals.