International Journal of Advances in Applied Sciences (IJAAS)
Vol. 13, No. 4, December 2024, pp. 834~843
ISSN: 2252-8814, DOI: 10.11591/ijaas.v13.i4.pp834-843

Batak Toba language-Indonesian machine translation with transfer learning using no language left behind

Cevin Samuel, Irsan Taufik Ali
Department of Electrical Engineering, Faculty of Engineering, Universitas Riau, Pekanbaru, Indonesia

Article Info

Article history:

Received Nov 16, 2023
Revised Jun 3, 2024
Accepted Jun 19, 2024

Keywords:

Batak Toba language
Low-resource languages
Neural machine translation
NLLB-200 model
sacreBLEU score
Transfer learning

ABSTRACT

This study focuses on neural machine translation (NMT) for low-resource languages (LRLs) pair, Batak Toba-Indonesian (bbc↔ind). The Batak Toba language is a critically endangered dialect of an Indonesian ethnic group, Batak. Recent advances in machine translation offer potential solutions, with transfer learning emerging as a promising approach for this language pair. We used a publicly available bbc↔ind parallel corpora from the Hugging Face datasets hub and employed the NLLB-200's distilled 600M variant model as the baseline model. Our models achieved sacreBLEU scores as follows: i) for bbc→ind, it achieved a score of 37.10 (+25.67, up from 11.43) and ii) for ind→bbc, it achieved a score of 30.84 (+25.82, up from 5.02). These results outperform all previous works in the task bbc↔ind machine translation and prove the validity of our approach.

Corresponding Author:

Cevin Samuel
Department of Electrical Engineering, Faculty of Engineering, Universitas Riau
Bina Widya Campus KM. 12.5, Pekanbaru 28293, Indonesia
Email: cevin.samuel3778@student.unri.ac.id

1. INTRODUCTION

Among the diverse linguistic landscape of Indonesia, home to an incredible array of languages and dialects, the Batak Toba language, spoken by millions in North Sumatra, is in trouble. Classified as endangered on the UNESCO list of endangered languages [1], the preservation of the language is of particular importance in the face of declining numbers of speakers and resources. An essential step towards maintaining such a language is opening doors for communication through the modern age, allowing the language to remain relevant and accessible to younger generations in the face of the dominant languages [2].

Cultural and historical significance, here are some significant features of its importance: i) Batak Toba is an essential component of the Batak people's cultural identity. It is utilized in traditional ceremonies, music, stories, and daily life, representing their distinct cultural legacy and history; ii) Batak Toba is an Austronesian language, contributing to Indonesia's and the world's linguistic variety. Losing it would devalue this unique linguistic heritage; and iii) Batak Toba provides insights into the Batak people's historical evolution and contacts with other cultures.

Potential impact of this work, there are several ways in which this research might contribute to the preservation and restoration of the language. The model can serve several purposes: i) promoting accessibility is by facilitating communication between Batak Toba speakers and Indonesian speakers, it can expand the resources and knowledge availability in both languages; ii) developing educational materials and resources in both Batak Toba and Indonesian, enhancing education for Batak Toba communities; iii) sustain the culture is the model can be integrated into other digital platforms and applications, further promoting Batak Toba; and, iv) conducting additional research, this study paves the way for future research in the Batak Toba language, encouraging scientists and developers to contribute to its advancement and preservation.

Transfer learning, a straightforward yet effective approach to improving model performance of low-resource neural machine translation (NMT), offers a promising approach for Batak Toba language restoration. By using existing pre-trained machine translation models like NLLB-200, we can accelerate the development of Batak Toba translation models, even with limited data. This approach can empower Batak Toba speakers to access information and connect with broader communities, ultimately contributing to the language's restoration efforts [3]. For instance, a person learning Batak Toba can utilize this model to find translations they need, making it a practical tool for language learning.

This paper fills this gap by utilizing transfer learning, a technique that is perfect for training models with limited data. We intend to: i) create a Batak Toba-Indonesian (bbc↔ind) machine translation model using transfer learning, demonstrating its feasibility for low-resource languages (LRLs); ii) evaluate the model's performance to assess its potential for real-world applications and future development; and iii) contribute to Batak Toba preservation by providing a tool for cultural, communication, and education promotion. By fulfilling these goals, we intend to pave the path for future study and development in Batak Toba machine translation, eventually contributing to the language's survival and resilience in the digital era.

2. RELATED WORK

2.1. Low-resource languages

Researchers have been actively working to improve the translation quality for LRLs, aiming to bridge the gap with high and medium-resource languages. One possible approach involves transferring knowledge from a high-resource language to the LRLs via a technique known as transfer learning [4]. This method uses current knowledge to increase translation accuracy in LRLs. Another important method is data augmentation. Unlike parallel data (where sentences exist in both languages), LRLs typically have much more readily available monolingual data (text in only the target language). This advantage allows for the creation of synthetic parallel data pairs, which have been shown to enhance translation accuracy [5].

For instance, a study focusing on Finno-Ugric languages demonstrated that training a multilingual model incorporating other languages within the same family led to significant improvements in Estonian translation quality [6]. It is notably shown in [7] that the closer the relationship between the high or medium-resource languages used in training and the LRLs, the greater the translation accuracy. While transfer learning has proven valuable for LRLs with limited parallel resources [8], the study also revealed instances where using an unrelated high-resource language resulted in even better translations.

2.2. NLLB-200

NLLB-200 is a large-scale multilingual machine translation model capable of translating between 202 languages, including many LRLs as identified in [9]. However, Batak Toba is one of the languages currently not supported. The model was trained on 18 billion parallel sentences covering over 40,000 translation directions. The training data was created using a combination of human-translated benchmarks (FLORES-200) and a new toxicity benchmark encompassing all FLORES-200 languages, ensuring translation quality and safety. As described in [10], the NLLB team employed bitext mining and back-translation techniques to construct the non-English-centric training dataset. Compared to previous state-of-the-art models like Deepnet [11] and M2M-100 [10], NLLB-200 achieves a 44% BLEU score improvement.

As reported in [9], by having extensive training data, broad language support, and promising evaluation results, the NLLB-200 model could serve as a strong foundation for training a bbc↔ind translation system. For the experiments in this study, we used the distilled 600M variant of the NLLB-200 model as the baseline for both our bilingual unidirectional models, i) bbc→ind and ii) ind→bbc. For the complete details on how we built and trained these models, please refer to the provided code here [12].

3. EXPERIMENTAL SETTINGS

3.1. Dataset and preprocessing

We used the NusaX dataset, a parallel corpus that contains bbc↔ind [13] language pair created by Indonesian natural language processing (IndoNLP). The IndoNLP team created the dataset [14] by translating SmSA [15], a recognized Indonesian sentiment analysis dataset that consists of reviews and comments from the IndoNLU benchmark [16]. Competent bilingual speakers were responsible for the translation. It went through refinement with human-assisted quality assurance. We refer the reader to [14] for more on the construction processes of the dataset.

At preprocessing time, we used the train and test subsets of the NusaX dataset as the training and evaluation datasets, respectively. Table 1 presents the information detailing the number of parallel sentences within the subsets we used. After loading up the dataset, aside from using the pre-trained built-in tokenizer

and data collator from the NLLB-200 model, we chose not to perform any additional normalization preprocessing (e.g., collapsing punctuation, removing special characters) onto the dataset.

Table 1. Number of parallel sentences in the (bbc↔ind) dataset

Subset	# Parallel sentences
train	6,606
test	2,000

While removing punctuation and special characters is a common preprocessing step in some NLP tasks, we opted to retain them in this instance for the following reasons: i) preserving potentially useful information: Punctuation marks can convey important information such as sentence endings, question types, and speaker emotions. Similarly, special characters like currency symbols or mathematical notations can hold specific meanings within the context. Removing these elements could potentially lead to misinterpretations during the translation process and ii) Domain-specific considerations: Batak Toba might utilize punctuation and special characters in unique ways essential for accurate translation. It is a relatively under-resourced language, so keeping these elements might be crucial to capturing the language nuances. It is important to note that the decision to maintain or remove punctuation and special characters varies based on the language pair, the translation model's intended application, and the dataset's features. Further research might be needed to discover the best preprocessing approach for subsequent iterations of this model.

The maximum sequence length is set to 64 tokens for both the encoder and decoder. This value is chosen based on the observed distribution of sentence lengths in the dataset. Setting the maximum sequence length too low could truncate informative parts of the sentences while setting it too high could lead to excessive computational cost and potential memory issues. Analyzing the dataset revealed that the vast majority of sentences fall well within the 64-token limit, striking a balance between capturing necessary information and computational efficiency.

Tables 2 and 3 show that the second column, text, in our dataset contains sentences written in ind, while the third column, the label contains sentences written in bbc. And since our approach in this study is to employ two completely separate and distinct bilingual unidirectional models, namely bbc→ind and ind→bbc; we simply switched the source and target languages. Specifically, for bbc→ind, we set the sentences from the text column as the target language, and conversely, for ind→bbc, we set it as the source language.

3.2. Implementation and evaluation

We used the Hugging Face Transformers library and the Google Colaboratory service platform for all of our training and implementation. Google Colaboratory is used to create, share, and run jupyter notebooks online, which are digital documents that can integrate executable code, images, and text in one file [17]. Google Colaboratory serves as a platform for running and sharing easily replicable and modifiable code, it is also widely utilized in machine learning research. Hugging Face transformers library is an open-source library that provides a comprehensive, latest transformer architecture through a unified API [18] to support a range of machine learning tasks. We then evaluate both our bilingual unidirectional models based on the sacreBLEU scores metric; an enhanced version of the BLEU metric designed to address the inconsistency in reporting scores from the BLEU metric. The main difference between sacreBLEU and BLEU is that sacreBLEU provides a standardized way of computing BLEU scores, making it easier to compare scores across different papers [19].

Table 2. Parallel sentence examples from the train subset

Id	Text	Label	Equivalent translation in English
0	*Hobi banget ngapain? selfie, sama mainan hp wkwk belajar juga sih tapi gak keseringan bgt wkw http://ask.fm/a/c0dmqfll*	*Hobi doma mangua? selfie, rap mar hp wkwk marsiajar juo sih tai inda na jotjot tu wkw http://ask.fm/a/c0dmqfll*	*You really like doing it what are you doing? selfie, and playing with my phone wkwk and studying too but not that frequently though wkw http://ask.fm/a/c0dmqfll*
1	*Thanks, babes, for coming @FNHaziqah_ @FtinAtighah @___szr. baju dh santek mai posing hancoq? #HasnolAnisSolemnization*	*Mauliate babes for coming @FNHaziqah_ @FtinAtighah @___szr. pahean nungga bagak dongan posing hancoq? #HasnolAnisSolemnization*	*Thanks, babes, for coming @FNHaziqah_ @FtinAtighah @___szr. my clothes were so good when I wanted to pose, they got ruined? #HasnolAnisSolemnization*
2	*Lumayan nyaman, cuma ac kurang dingin dan kebersihan perlu ditingkatkan*	*Menak do nian, alai ac dang ngali jala haison porlu ditingkathon*	*It's quite comfortable, but the ac isn't cold enough and the cleanliness needs to be improved*

Table 3. Parallel sentence examples from the test subset

Id	Text	Label	Equivalent translation in English
0	*Receptionist kurang ramah, lift kotor dan bau, air kamar mandi kurang lancar tapi air panas dari shower mantab, kebersihan bantal dan sprei perlu dievaluasi lagi. Amat disayangkan potensi yg sangat baik tidak didukung profesionalisme SDM nya, fungsi kontrolnya perlu ditingkatkan*	*Revepsionis hurang rama, lift dursun rap bau, aek kamar maridi hurang lancar tai aek milas ngen paccur mantab, kaiasan battal rap sipere porlu dievaluasi buse. amana disayangkon potensi na sattak burju inda didukung profesionalisme SDM na, fungsi kontrolna porlu di tingkatkon*	The receptionist is not friendly, the elevator is dirty and smelly, the bathroom water is not smooth but the hot water from the shower is good, the cleanliness of pillows and bed linen needs to be evaluated again. It is a pity that the excellent potential is not supported by the professionalism of its human resources, the control function needs to be improved
1	*Biar pulpen itu menjadi kenangan dan ingatan bahwa saya pernah datang untuk bertamu dan bercerita hingga menulis. senyum yang tidak akan pernah hilang dan tawa yang selalu akan saya ingat, terima kasih. mari kita menutup hal itu dengan lagu ini: [URL]*	*Anso pulpen i manjadi kenangan rap ingotan bahwa au ujung ro giot martamu rap marcarito sampe manulis, mikim na inda nakkan unjung mago rap tata na salalu au ingot, tarimo kasih. mari hita manutup hal i rap lagu on: [URL]*	Let that pen be a memory and a reminder that I once came to visit and tell stories to write. a smile that will never disappear and a laugh that I will always remember, thank you. let's close it with this song: [URL]
2	*Kartosoewiryo itu beragama dan mau bunuh Soekarno. Yang bunuh Sayyidina Ali ketika solat subuh itu seorang Hafidz Quran, Ibnu Muljam. Sampai disini sudah jelas belum, bro? [URL]*	*Kartosoewiryo maragama rap giot mambunuh soekarno.na mambunuh sayyidina Ali hatia sumbayang subuh i seorang Hafidz Quran, ibnu muljam. sampe dison madung jelas sanga napedo, bro? [URL]*	Kartosoewiryo was religious and wanted to kill Soekarno. The one who killed Sayyidina Ali during the dawn prayer was a Quran Hafidz, Ibn Muljam. Is it clear yet, bro? [URL]

4. RESULTS AND DISCUSSION

4.1. Zero-shot capabilities of NLLB-200

We first assessed the baseline performance of the NLLB-200 model without any training. This involved measuring its evaluation loss and sacreBLEU scores on the test subset in Table 4, followed by an evaluation of its zero-shot translation capabilities for bbc→ind and ind→bbc language pairs (Tables 5 and 6) Our findings upon looking at the low sacreBLEU scores (0.0 to 12.44) and code-mixed translations (English/eng instead of bbc) in Tables 5 and 6 indicate that employing a pre-trained language model, particularly the NLLB-200 model, is not a usable strategy for zero-shot translation in the case of new, unseen language. They also highlight that the absence of bbc labels in the NLLB-200 model poses a challenge, causing the model to default to eng when bbc is not recognized as a label option in the source/target language. Even when bbc is set as an option, the model fails to recognize it and defaults to eng.

Consequently, in the bbc→ind direction, where the source sentences are filled with unrecognized eng text (which is actually in bbc), they cannot align with the ind corpus, causing the majority of the predictions to remain in bbc (Table 5). On the other hand, in the ind→bbc direction, where bbc is not recognized as a target label, the task becomes ind→eng direction, resulting in all predictions being in eng (Table 6). These findings are in line with [20], where it is noted that while large pre-trained language models can generate correct translations in zero-shot scenarios, they are also prone to producing translations in the wrong language.

4.2. Training

We leverage the NLLB-200 model for bbc↔ind translation through transfer learning. From Table 4, we can depict from the high evaluation loss as well as the low sacreBLEU scores that the NLLB-200 model needs further tuning or adjustments to achieve optimal performance. Here, we detail the selected hyperparameters and their justifications during training:

i) Batch size

We employ an initial batch size of 4 for the training subset and 16 for the test subset. These values are chosen based on the available GPU memory on the employed platform (NVIDIA Tesla T4 32 GB). Smaller batch sizes allow for more frequent parameter updates during training, potentially leading to faster convergence. However, excessively small batches can also increase training time due to the overhead associated with processing individual batches. The chosen batch sizes represent a compromise between these considerations, allowing for efficient utilization of the available memory while maintaining reasonable training speed.

ii) Optimizer

We apply the AdamW (Adam with weight decay [21]) optimizer with specific hyperparameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$, $weight\_decay = 0.3$. AdamW is a widely used optimizer in deep learning known for its efficiency and stability in handling large models. The chosen hyperparameter values are based

on best practices and empirical findings from previous studies on fine-tuning pre-trained models for machine translation tasks.

iii) Weight decay

We apply weight decay to all parameters except the bias and LayerNorm weights with a value of 0.3. This technique helps in preventing overfitting by penalizing large parameter values during training. The chosen value balances the need for regularization to avoid overfitting and the need to allow the model to learn effectively.

iv) Learning rate scheduler

We adopt a linear learning rate scheduler with a warmup phase (Figure 1), as implemented in [22]. The initial learning rate is set to 5×10^{-5}. The warmup phase lasts for 825 steps, where the learning rate increases linearly from 0 to 5×10^{-5}. For the remaining steps (7,425), the learning rate linearly decreases from 5×10^{-5} to 0. This approach allows the model to initially explore the parameter space more freely, followed by a gradual decrease in the learning rate to fine-tune the parameters for optimal performance.

v) Mixed precision training

To optimize training efficiency and reduce memory usage, we employ mixed precision training with FP16 as implemented in the Fairseq library [23]. This technique allows the model to perform computations using a lower precision format (16-bit) while maintaining the gradients in a higher precision format (32-bit). This approach reduces memory usage without reducing the model's accuracy.

vi) Total training time

The training process took approximately 1 hour, 51 minutes for the bbc→ind direction, and 2 hours, 21 minutes for the ind→bbc direction. These durations are influenced by factors like the dataset size, chosen hyperparameters, and available hardware resources. The training loss and sacreBLEU u scores graphs for our two trainings are presented in Figure 2, especially in Figures 2(a) and (b).

Table 4. Evaluation loss and sacreBLEU scores of the model on the test subset before training

Direction	Evaluation loss (lower is better)	SacreBLEU (higher is better)
bbc→ind	7.95	11.43
ind→bbc	8.37	5.02

Table 5. Evaluation of zero-shot capability in bbc→ind direction

No	Source	Target	Prediction	SacreBLEU (higher is better)	Equivalent translation of target in English
1	*Adong na boi hu urupi?*	*Ada yang bisa aku bantu?*	*Apakah Anda takut dengan organisasi?*	8.12	Is there anything I can help you with?
2	*Ubat ni sude sahit roha na sonang.*	*Obat segala penyakit adalah hati yang gembira.*	*Ubat adalah bagian selatan dari sahit roha dan sonang.*	4.99	The cure for all diseases is a joyful heart.
3	*Teratur mangan dohot cukup parmodom mu dah, asa unang marsahit be muse.*	*Makan teratur dan istirahat yang cukup ya, biar ga sakit lagi.*	*Teratur mangan dohot cukup parmodom mu dah, asa unang marsahit menjadi muse.*	3.74	Eat regularly and get enough rest, so you don't get sick again.
4	*Jolma na tarbereng sonang dijoloni natorop, dang tentu jolma naso adong masalahna, alai ima jolma na tangis disihabunian.*	*Orang yang bahagia di depan orang banyak, bukanlah orang yang tidak punya masalah, tetapi dialah orang yang menangis dikala orang tidak ada.*	*Jolma dan tarbereng sonang dijoloni natorop, dang tentu jolma naso adong masalahna, alai ima jolma dan tangis disihabunian.*	1.98	The person who is happy in front of people is not the one who has no problems, but he is the one who cries when people are not around.
5	*Ale Tuhan, lean ma hasabaron tu keluarga na dibagas parmaraan. Alana takkas do huboto nami, sude na marhosa dohot na manggulmit di portibi on to saat tujoloMu do ujungna. Anggiat ma Ho ale Jahowa manjalo toddi ni keluarga nami on, sahat ro di tanganMu ma ibana, jala pahuddul ma i di inganan na sonang.*	*Ya Tuhan, tabahkanlah keluarga yang sedang dilanda musibah. Karena, semua makhluk di dunia pasti akan ada saatnya menemui Sang Pencipta. Semoga Engkau menerima arwah saudara kami dengan tanganMu yang terbuka lebar menyambutnya dan menempatkannya di tempat yang terindah.*	*Ya Tuhan, lean ma hasabaron tu keluarga dan dibagas parmaraan. Alana takkas do huboto denganku, selatan dan marhosa dohot dan manggulmit di portibi pada saat itu aku pergi ke mu akhirna. Anggiat ma ho ale jehova dan keluarga ku pada saat ini, sahat kami berada di tanganMu dan ibana, jala pahuddul dan di sini sama dengan sonang.*	3.10	O God, be strong for the family that is being hit by disaster. Because, all creatures in the world will definitely have their time to meet the Creator. May You accept our brother's spirit with Your arms wide open to welcome him and place him in the most beautiful place.

Table 6. Evaluation of zero-shot capability in ind→bbc direction

No	Source	Target	Prediction	SacreBLEU (higher is better)	Equivalent translation of target in English
1	*Turun disini lah kita.*	*Turun dison ma hita.*	Get down here.	12.44	Get down here.
2	*Ini pengalaman pertamaku.*	*On ma pengalamanku na parjolo.*	This is my first experience.	8.12	This is my first experience.
3	*Jika kamu pergi, hatiku akan pergi bersamamu.*	*Molo lao ho, rohakku dohot mai lao.*	If you go, my heart will go with you.	4.46	If you go, my heart will go with you.
4	*Ketika menghadapi masalah apapun dalam kehidupan ini maka terimalah dengan hati yang sabar dan iman yang kuat.*	*Bahenma dirimu songon laut naluas, manang songon dia pe masalah naroh tungolumu, jalo ma dohot roha naserep dohot iman na gogo.*	N/A	0.0	When facing any problem in life, accept it with a patient heart and strong faith.
5	*Biarlah orang lain memuji engkau dan bukan mulutmu, orang yang tidak kau kenal dan bukan bibirmu sendiri. Batu adalah berat dan pasir pun ada beratnya, tetapi lebih berat dari kedua-duanya adalah sakit hati terhadap orang bodoh.*	*Halak na asing tagonan mamuji ho, unang tung pamanganmu sandiri, halak sileban tagonan, unang bibirmu sandiri. Dokdok do batu, jala borat horsik, alai dumokdok sian duansa do anggo hamurhingon ni halak na oto.*	Let another praise thee, and not thy mouth; a stranger, and not thy own lips. The stone is heavy, and the sand heavy; but more than these is the sorrow of the fool.	1.37	Let another praise thee, and not thy mouth; a stranger, and not thy own lips. The stone is heavy, and the sand heavy; but more than these is the sorrow of the fool.
6	*Jaman dahulu hiduplah seorang raja yang Bernama Raja Rahat yang berkuasa di Samosir. Raja Rahat adalah raja yang dikenal rakyatnya sebagai raja yang bijaksana dan adil. Para rakyatnya pun senang memiliki raja seperti Raja Rahat. Raja Rahat juga memiliki permaisuri yang baik hati dan memiliki anak satu yang beri nama Manggale.*	*Na jolo adong ma raja na margoar Raja Rahat na marhuta i Samosir. Raja Rahattakkas do i tanda akka raiatna raja nabisuk jala natigor. Sonang do akka raiatna mangida raja Rahat on. Raja Rahat on mangarihon parsonduk bolon na tong marbasa roha jala mangarihon sada ianakhon na margoar Manggale.*	N/A	0.0	Long ago there lived a king named King Rahat who ruled in Samosir. King Rahat was known by his people as a wise and just king. His people were happy to have a king like King Rahat. King Rahat also had a queen who was kind and had one child named Manggale.

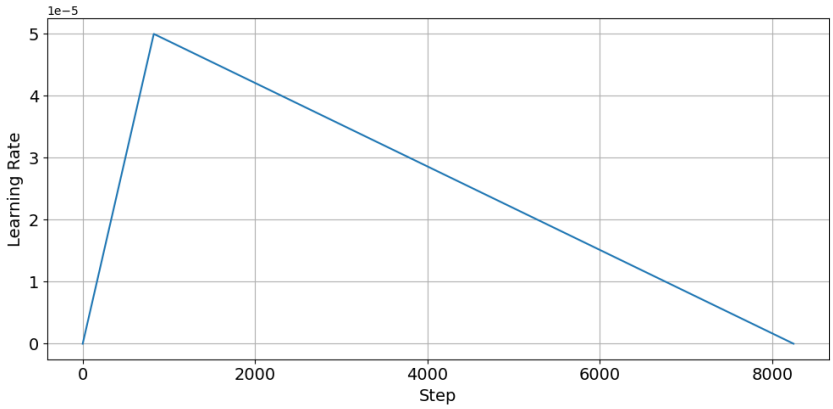
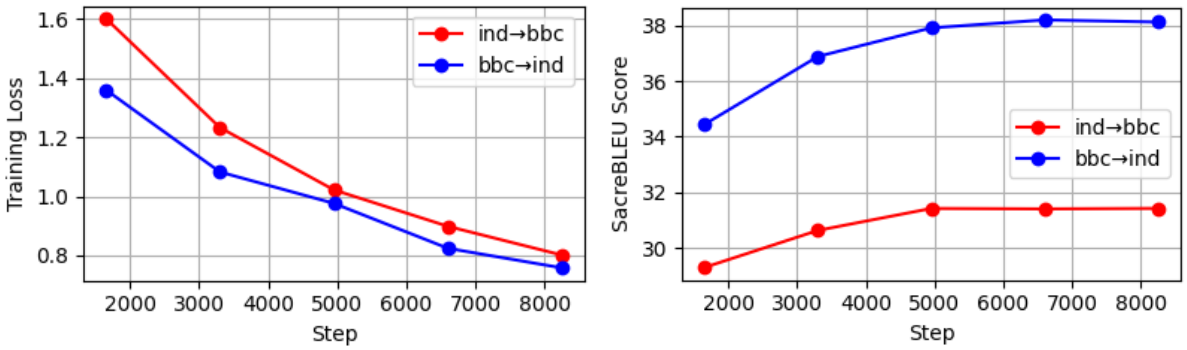


Figure 1. Learning rate schedule with warmup phase for training the NLLB-200 model on the bbc↔ind dataset

4.3. Results

In Figure 2, we illustrate the results achieved by training the NLLB-200 model on LRLs parallel corpora for both ind→bbc and bbc→ind language directions. Our analysis of these findings suggests that leveraging LRLs parallel corpora for pre-trained language model training is a promising approach, even for scenarios involving new, unseen languages, Figure 2(a) shows Training loss and Figure 2(b) shows sacreBLEU scores. This approach can alleviate the limitations associated with the scarcity of data in LRLs. In comparison to the sacreBLEU scores of the model before training (Table 4), the model demonstrated a significant improvement, achieving sacreBLEU scores of up to 37.10 (+25.67, up from 11.43) in bbc→ind and 30.84 (+25.82, up from 5.02) in ind→bbc.



(a) (b)

Figure 2. Model after training in ind→bbc and bbc→ind directions (lower training loss is better, higher sacreBLEU score is better) of (a) Training loss and (b) sacreBLEU scores

The results demonstrate that NLLB-200 is suited to producing high-quality translations, even for a new and unseen LRLs pair. This capability is demonstrated by its proficiency in training with only 6,600 parallel sentences from the training subset, which were previously new and unseen by the model before training. For more examples, we can compare the sacreBLEU scores on the test subset in Tables 5 and 6. After the training in Tables 7 and 8, the improvement in sacreBLEU scores ranges from +0.63 (Parallel sentence example no. 2 of Tables 6 and 8) up to +46.78 points (Parallel sentence example no. 3 of Tables 6 and 8).

Table 7. Evaluation of the model after training in bbc→ind direction

No	Source	Target	Prediction	SacreBLEU (higher is better)	Equivalent translation of target in English
1	*Adong na boi hu urupi?*	*Ada yang bisa aku bantu?*	*Ada yang bisa saya bantu?*	37.99	Is there anything I can help you with?
2	*Ubat ni sude sahit roha na sonang.*	*Obat segala penyakit adalah hati yang gembira.*	*Obat semua sakit hati yg bahagia.*	6.77	The cure for all diseases is a joyful heart.
3	*Teratur mangan dohot cukup parmodom mu dah, asa unang marsahit be muse.*	*Makan teratur dan istirahat yang cukup ya, biar ga sakit lagi.*	*Teratur makan dan cukup tidur ya, biar ga sakit lagi.*	50.52	Eat regularly and get enough rest, so you don't get sick again.
4	*Jolma na tarbereng sonang dijoloni natorop, dang tentu jolma naso adong masalahna, alai ima jolma na tangis disihabunian.*	*Orang yang bahagia di depan orang banyak, bukanlah orang yang tidak punya masalah, tetapi dialah orang yang menangis dikala orang tidak ada.*	*Orang yang terlihat bahagia di kampung halaman, bukan tentu orang yang tidak punya masalah, tapi adalah orang yang menangis disebuah rumah.*	28.63	The person who is happy in front of people is not the one who has no problems, but he is the one who cries when people are not around.
5	*Ale Tuhan, lean ma hasabaron tu keluarga na dibagas parmaraan. Alana takkas do huboto nami, sude na marhosa dohot na manggulmit di portibi on to saat tujoloMu do ujungna. Anggiat ma Ho ale Jahowa manjalo toddi ni keluarga nami on, sahat ro di tanganMu ma ibana, jala pahuddul ma i di inganan na sonang.*	*Ya Tuhan, tabahkanlah keluarga yang sedang dilanda musibah. Karena, semua makhluk di dunia pasti akan ada saatnya menemui Sang Pencipta. Semoga Engkau menerima arwah saudara kami dengan tanganMu yang terbuka lebar menyambutnya dan menempatkannya di tempat yang terindah.*	*Ya Tuhan, berilah kesabaran kepada keluarga yang berada di dalam surga. Karena kami pasti tahu, bahwa semua yang berdosa dan mengikhlaskan diri di dunia ini akan datang menjelangMu pada akhirnya. Semoga Engkau tetapi juga Ya Allah menerima seluruh keluarga kami, sampai mereka sampai di tanganMu, dan mendudukkan mereka di tempat yang menyenangkan.*	7.46	O God, be strong for the family that is being hit by disaster. Because, all creatures in the world will definitely have their time to meet the Creator. May You accept our brother's spirit with Your arms wide open to welcome him and place him in the most beautiful place.

Table 8. Evaluation of the model after training in ind→bbc direction

No	Source	Target	Prediction	SacreBLEU (higher is better)	Equivalent translation of Target in English
1	*Turun disini lah kita.*	*Turun dison ma hita.*	*Turun tuson ma hita.*	42.73	Get down here.
2	*Ini pengalaman pertamaku.*	*On ma pengalamanku na parjolo.*	*on pengalaman pertama ku.*	8.75	This is my first experience.
3	*Jika kamu pergi, hatiku akan pergi bersamamu.*	*Molo lao ho, rohakku dohot mai lao.*	*Molo lao ho, rohaku laho rap dohot ho.*	29.07	If you go, my heart will go with you.
4	*Ketika menghadapi masalah apapun dalam kehidupan ini maka terimalah dengan hati yang sabar dan iman yang kuat.*	*Bahenma dirimu songon laut naluas, manang songon dia pe masalah naroh tungolumu, jalo ma dohot roha naserep dohot iman na gogo.*	*Hatia mangadopi masalah aha pe dalam hidup on maka terima ma rap hati na sabar rap iman na gogo.*	11.29	When facing any problem in life, accept it with a patient heart and strong faith.
5	*Biarlah orang lain memuji engkau dan bukan mulutmu, orang yang tidak kau kenal dan bukan bibirmu sendiri. Batu adalah berat dan pasir pun ada beratnya, tetapi lebih berat dari kedua-duanya adalah sakit hati terhadap orang bodoh.*	*Halak na asing tagonan mamuji ho, unang tung pamanganmu sandiri, halak sileban tagonan, unang bibirmu sandiri. Dokdok do batu, jala borat horsik, alai dumokdok sian duansa do anggo hamurhingon ni halak na oto.*	*Asa di puji ho jolma na asing jala dang di sungkun mu, jolma na dang di boto ho jala dang di bibirmu sandiri. Batu na borat jala pasir pe adong boratna, alai lobi borat sian nadua-dua i ima hatcit roha tu jolma oto.*	5.44	Let another praise thee, and not thy mouth; a stranger, and not thy own lips. The stone is heavy, and the sand heavy; but more than these is the sorrow of the fool.
6	*Jaman dahulu hiduplah seorang raja yang Bernama Raja Rahat yang berkuasa di Samosir. Raja Rahat adalah raja yang dikenal rakyatnya sebagai raja yang bijaksana dan adil. Para rakyatnya pun senang memiliki raja seperti Raja Rahat. Raja Rahat juga memiliki permaisuri yang baik hati dan memiliki anak satu yang beri nama Manggale.*	*Na jolo adong ma raja na margoar Raja Rahat na marhuta i Samosir. Raja Rahattakkas do i tanda akka raiatna raja nabisuk jala natigor. Sonang do akka raiatna mangida raja Rahat on. Raja Rahat on mangarihon parsonduk bolon na tong marbasa roha jala mangarihon sada ianakhon na margoar Manggale.*	*Zaman najolo mangolu ma sada raja na margoar Raja Rahat na markuasa di Samosir. Raja Rahat ima raja na di boto rakyatna sebagai raja na bijak jala adil. Para rakyatna pe sonang puna raja songon Raja Rahat. Raja Rahat pe puna permaisuri na burju roha jala puna anak sada na mangalean goar Manggale.*	15.12	Long ago there lived a king named King Rahat who ruled in Samosir. King Rahat was known by his people as a wise and just king. His people were happy to have a king like King Rahat. King Rahat also had a queen who was kind and had one child named Manggale.

4.4. Benchmarking our results against prior state-of-the-art Batak Toba-Indonesian translation

One of our most important achievements is that, in comparison to previous studies on bbc↔ind machine translation, specifically referencing [14], [24], our approach achieves the best result in sacreBLEU scores. As seen in Table 9, our models have a sacreBLEU of i) 37.10 for bbc→ind and ii) 30.84 for ind→bbc; both are higher compared to [14]'s sacreBLEU of i) 20.94 for bbc→ind and ii) 18.41 for ind→bbc; [24]'s sacreBLEU of i) 20.48 for bbc→ind and ii) 18.41 for ind→bbc. It's worth noting that while [24] employed the same baseline model as ours the NLLB-200's distilled 600M variant model in one of their experiments, they didn't introduce it to the bbc↔ind language pair. This limitation arises from the fact that the NLLB-200 model is not trained on the bbc↔ind language pair [9], which is precisely the focus of our work. To view the results as live models, access these: ind→bbc [25], bbc→ind [26]. The code including the implementation details can be found at [12].

Table 9. Comparison of our model in bbc↔ind with previous studies

Study	bbc→ind SacreBLEU (higher is better)	ind→bbc SacreBLEU (higher is better)	The baseline model used which achieves the highest SacreBLEU (bbc→ind)	The baseline model used which achieves the highest SacreBLEU (ind→bbc)
[14]	20.94	18.41	PBSMT	IndoBARTv2
[24]	20.48	18.41	mT5	IndoBARTv2
Ours	**37.10**	**30.84**	NLLB-200	NLLB-200

5. CONCLUSION

Based on our findings, using a large pre-trained language model like the NLLB-200 as a baseline proves to be effective for LRLs translation, specifically for bbc↔ind pairs, even for entirely unseen languages. For example, some of our evaluations show impressive gains in sacreBLEU scores, reaching

+46.78 (bbc→ind) and +30.29 (ind→bbc). However, translating to/from entirely unseen languages in zero-shot settings revealed a crucial limitation of the NLLB-200 model: generating correct translations in the wrong language. While pre-trained models hold promise for zero-shot translation, challenges remain, especially for unseen languages and domains. Further research is necessary to overcome these hurdles and enhance pre-trained model effectiveness in zero-shot translation.

REFERENCES

[1] UNESCO WAL, "Batak Toba," UNESCO. Accessed: Feb. 25, 2024. [Online]. Available: https://en.wal.unesco.org/languages/batak-toba

[2] H. V. S, Sumarsih, and R. Husein, "Toba Batak language shift in Rantau Selatan," in *Proceedings of The 4th Annual International Seminar on Transformative Education and Educational Leadership (AISTEEL)*, Medan: EPrints, 2019.

[3] E. Sinambela and L. L. Hutagaol, "Idiomatic translation of umpasa in delivering ulos in Toba Batak wedding ceremony," *Explora*, vol. 9, no. 1, pp. 67–87, Apr. 2023, doi: 10.51622/explora.v9i1.1243.

[4] J. Gu, H. Hassan, J. Devlin, and V. O. K. Li, "Universal neural machine translation for extremely low resource languages," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 344–354. doi: 10.18653/v1/N18-1032.

[5] R. Sennrich and B. Zhang, "Revisiting low-resource neural machine translation: a case study," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 211–221. doi: 10.18653/v1/P19-1021.

[6] M. Tars, T. Purason, and A. Tättar, "Teaching unseen low-resource languages to large translation models," in *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 375–380.

[7] M. Tars, A. Tättar, and M. Fišel, "Extremely low-resource machine translation for closely related languages," in *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, Reykjavik, Iceland: Linköping University Electronic Press, Sweden, 2021, pp. 41–52.

[8] T. Kocmi and O. Bojar, "Trivial transfer learning for low-resource neural machine translation," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 244–252. doi: 10.18653/v1/W18-6325.

[9] NLLB Team *et al.*, "No language left behind: scaling human-centered machine translation," Jul. 2022, [Online]. Available: http://arxiv.org/abs/2207.04672

[10] A. Fan *et al.*, "Beyond English-centric multilingual machine translation," *Journal of Machine Learning Research*, vol. 22, pp. 1–48, 2021.

[11] H. Wang, S. Ma, L. Dong, S. Huang, D. Zhang, and F. Wei, "DeepNet: scaling transformers to 1,000 Layers," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, pp. 1–14, doi: 10.1109/TPAMI.2024.3386927.

[12] GitHub, "Batak Toba language-Indonesian machine translation with transfer learning using No Language Left Behind," GitHub, Inc. Accessed: Feb. 25, 2024. [Online]. Available: https://github.com/caffeineeee/batak_toba_indonesian_nmt

[13] Hugging Face, "Datasets: indonlp/nusatranslation_mt," Hugging Face. Accessed: Feb. 24, 2024. [Online]. Available: https://huggingface.co/datasets/indonlp/nusatranslation_mt

[14] G. I. Winata *et al.*, "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*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 815–834. doi: 10.18653/v1/2023.eacl-main.57.

[15] A. Purwarianti and I. A. P. A. Crisdayanti, "Improving Bi-LSTM performance for Indonesian sentiment analysis using paragraph vector," in *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, IEEE, Sep. 2019, pp. 1–5. doi: 10.1109/ICAICTA.2019.8904199.

[16] B. Wilie *et al.*, "IndoNLU: benchmark and resources for evaluating Indonesian natural language understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Suzhou, China: Association for Computational Linguistics, 2020, pp. 843–857.

[17] L. C. Timpe, "An online Google Colab project for exploring the SARS CoV-2 genome and mRNA vaccines," *Biochemistry and Molecular Biology Education*, vol. 51, no. 2, pp. 209–211, Mar. 2023, doi: 10.1002/bmb.21711.

[18] T. Wolf *et al.*, "Transformers: state-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.

[19] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 186–191. doi: 10.18653/v1/W18-6319.

[20] K. Gupta, "MALM: mixing augmented language modeling for zero-shot machine translation," in *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, Taipei, Taiwan: Association for Computational Linguistics, 2022, pp. 53–58.

[21] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," Nov. 2017, [Online]. Available: http://arxiv.org/abs/1711.05101.

[22] E. A. Platanios, O. Stretcu, G. Neubig, B. Poczos, and T. M. Mitchell, "Competence-based curriculum learning for neural machine translation," Mar. 2019, [Online]. Available: http://arxiv.org/abs/1903.09848

[23] S. Samsi, M. Jones, and M. M. Veillette, "Compute, time and energy characterization of encoder-decoder networks with automatic mixed precision training," in *2020 IEEE High Performance Extreme Computing Conference (HPEC)*, IEEE, Sep. 2020, pp. 1–6. doi: 10.1109/HPEC43674.2020.9286241.

[24] W. Wongso, A. Joyoadikusumo, B. S. Buana, and D. Suhartono, "Many-to-many multilingual translation model for languages of Indonesia," *IEEE Access*, vol. 11, pp. 91385–91397, 2023, doi: 10.1109/ACCESS.2023.3308818.

[25] C. Samuel, "indonesian_to_bataknese_translation_model," Hugging Face. Accessed: Feb. 24, 2024. [Online]. Available: https://huggingface.co/kepinsam/ind-to-bbc-nmt-v1

[26] C. Samuel, "bataknese_to_indonesian_translation_model," Hugging Face. Accessed: Feb. 24, 2024. [Online]. Available: https://huggingface.co/kepinsam/bbc-to-ind-nmt-v1

BIOGRAPHIES OF AUTHORS

Cevin Samuel is an Informatics undergraduate student from Department of Electrical Engineering, Faculty of Engineering, Universitas Riau, Pekanbaru, Indonesia. He is skilled in full-stack web development and has a research interest in deep learning, particularly language models. He can be contacted at email: cevin.samuel3778@student.unri.ac.id and cevin.samuel@yahoo.com.

Irsan Taufik Ali is a lecturer at the Department of Electrical Engineering, Faculty of Engineering, Universitas Riau, Pekanbaru, Indonesia. He obtained his Doctorate in Informatics from Universitas Indonesia (UI), Indonesia, Master's degree in Information Technology from Universitas Gadjah Mada (UGM), Indonesia, and Bachelor's degree in Informatics from Universitas Islam Indonesia (UII), Indonesia. His research interest includes machine learning, deep learning, and information technology. He can be contacted at email: irsan.ali@lecturer.unri.ac.id.