Multilingual Translation for Indonesian Languages

Aditi Sharma, Sidharth Kathpal and Venny Ayudiani
Carnegie Mellon University
apsharma,skathpal,vayudian@andrew.cmu.edu

Abstract

Indonesia is the one of the most linguistically diverse countries, second only to Papua New Guinea. However, there is a scarcity in the amount of Natural Language Processing (NLP) research done for Indonesian languages, as discussed in (Aji et al., 2022). Figure 1 illustrates this challenge. In addition, there is a shortage of the amount of data available for these languages. This motivates the need to explore and improve the state-of-the-art of Machine Translation for Indonesian languages. The task of machine translation is important especially for the generation of data from low resource languages text.

1 Introduction

Machine Translation is a subfield of NLP that investigates the use of software to translate text or speech from one language to another. It is a rapidly-growing field that has seen great progress in the last few years with the advent of deep learning models, achieving state-of-the-art results in many high-resource language pairs. However, the progress for low-resource languages has lagged behind due to the limited availability of parallel data.

In this project, we look at the problem of translation in Indonesian languages, specifically those belonging to the Western Malayo-Polynesian region. We leveraged choosing similar transfer languages in order to perform multilingual training and data augmentation through backtranslation to improve multilingual transfer for those low-resource languages.

We also looked at two basic deep learning frameworks - OpenNMT (Klein et al., 2017) and Fairseq (Ott et al., 2019) to setup a few initial baselines to base our results off. The model from OpenNMT that we used was the pyramidal deep bidirectional encoder, and the fairseq architecture was based on the WMT20 task models based on the transformer medium. The idea behind implementing these pipelines was to actually discern and distinguish which would perform better and to add diversity in terms of the types of models that were used for producing baseline information using techniques which have become more prevelant in the machine translation domain in the last couple of years.

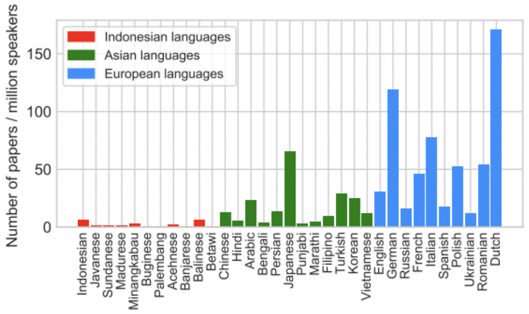

Figure 1: Research papers per million speakers (Aji et al., 2022)

2 Research Questions

2.1 Existing Work

Machine translation has been widely studied for decades, with various rule-based and corpus-based approaches. There are three prevalent rule-based approaches - direct, transfer and interlingua.

Systems that adopt a direct approach involve a word-for-word translation, without any intermediate representation. Direct rules are easier to learn automatically and such systems leverage structural similarities between languages. However, dealing with highly inflected languages becomes challenging and large systems become unmanageable. Further, no knowledge can be reused for new language pairs.

Transfer-based machine translation consists of multiple stages. First, the source text is analysed and transformed into abstract representations. Next, these representations are mapped to corresponding representations that are oriented to the target language. Finally, these are used to generate the target text. Transfer-based approaches leverage both the resource efficiency of interlingua systems and the simple implementation of direct systems.

Finally, the interlingua refers to an a universal language-independent meaning that the source text is transformed into, which in turn is used to generate the corresponding target text.

2.2 Our Contributions

The intended users of our work are researchers in the NLP community, specifically those focusing on machine translation for Indonesian languages. We explore various transfer-based machine translation approaches and build an end-to-end translation pipeline for machine translation in Indonesian languages.

Further, we extend the implementation of IndoBART to support 'Minangkabau' - a language for which there exists very minimal machine translation research. We also explore data augmentation techniques to boost the performance of our models by expanding the vocabulary of the corpora. As far as we are aware, this is the first research work that uses backtranslation to improve machine translation for Indonesian languages using the FLORES-101 dataset. All our work can be found publicly available in the following repository[1].

3 Data

Since Indonesia is a linguistically rich country with over 700 languages spoken across the country, it would be better to consider relatively similar languages that are interchangeable to a large extent. For this reason, we considered the following languages (and the corresponding language codes that we use throughout this report):

- Indonesian (ind)
- Javanese (jav)
- Sundanese (sun)
- Minangkabau (min)

These are low-resource languages and the most easily available dataset for most of these languages is the Bible dataset.

More details about the datasets we have used for training our models can be found in Table 1. In addition, we have used the WMT21 Shared Task FLORES-101 dataset (Goyal et al., 2021) for evaluation and conducted an ablation study with varying number of sentences. The FLORES-101 dataset is publicly available dataset with multiple languages which was aimed at breaking down language barriers. The dataset mainly incorporated a high amount of low resource languages, one of which was 'Javanese', which is the one we have used for evaluation. Around 80% of the languages in the dataset are low resource languages. The reason for only capturing the 'Javanese' and 'Indonesian' translation task from this particular dataset was because the dataset didn't contain the other two languages we are currently working on.

Lang	Dataset	Train	Dev	Test
jav-ind	Bible	5968	797	1193
sun-ind	Bible	5967	797	1193
min-ind	Wikipedia	11571	1600	3200
jav-ind	FLORES-101	19595	997	1012

Table 1: Corpus size used for Machine Translation

4 Models

4.1 Baselines

The baseline results for 'Javanese' and 'Sundanese' were taken from the machine translation task in (Cahyawijaya et al., 2021).

For 'Minangkabau', the baselines were taken from (Koto and Koto, 2020). They adopted a word-to-word (W2W) translation approach using bilingual dictionary. A manual evaluation by 2 'Indonesian' and 'Minangkabau' speakers showed that this W2W approach was significantly better than LSTM in terms of adequacy and similar in terms of fluency. In 2, it can be observed that the baseline BLEU scores for these translations are very high. This is because of the high similarity between 'Minangkabau' and 'Indonesian', which have overlaps of lexicons and syntax. In particular, 'Minangkabau' and 'Indonesian' generally have similar word and character lengths, which is why the W2W model is sufficient to achieve high BLEU scores for these translations.

4.2 Techniques Implemented

Initially we wanted to create our own baselines using 2 approaches - a seq-to-seq model and a transformer-based architecture. Beyond that we also look into using a pre-trained language model

[1] https://github.com/sidharthkathpal/IndoMultilingualMT.git

Language	Baseline	OpenNMT	Fairseq	IndoBART	Multilingual Training
jav-ind	34.20	21.35	24.1	38.71	**44.88**
ind-jav	26.06	16.13	20.44	36.10	**44.35**
sun-ind	16.11	3.49	11.41	20.95	**23.13**
ind-sun	12.40	3.18	7.92	15.48	**23.37**
min-ind	64.54	18.86	67.08	**74.65**	72.89
ind-min	55.08	17.52	53.43	**65.39**	64.14

Table 2: BLEU scores for different techniques explored. The baseline for 'Javanese' and 'Sundanese' were taken from (Cahyawijaya et al., 2021) and baseline for 'Minangkabau' were taken from (Koto and Koto, 2020).

which was trained on monolingual corpora for 'Sundanese', 'Javanese', and 'Indonesian'. The idea was to leverage this pretrained model to improve the machine translation task. The eventual model we settled upon was trained as a multilingual machine translation model which incorporated the dictionaries from all the languages mentioned above. We eventually fine-tuned these for the parallel translation tasks.

4.2.1 OpenNMT

The OpenNMT is a open source machine translation toolkit developed in 2017 by a team from Harvard University, which prioritizes the efficiency and modularity aimed at supporting machine translation research. The system is basically a complete library which not only consists of a vanilla NMT models but also supports for attention, gating, stacking, input feeding, regularization, beam search and all other options necessary for state-of-the-art performance.

4.2.2 Fairseq

Fairseq is a widely-used sequence-modeling toolkit that can train custom models for a variety of NLP tasks such as translation and summarization. We trained a transformer translation model for all our language pairs as an alternative baseline model. The main aim for the researchers building this was to help progress custom training of models for the neural machine translation task. The toolkit is based off of pytorch to make sure the content used in the toolkit is generalisable and this toolkit provides you with the models used in the shared machine translation task. Which is why we used this as the baseline as it presents us with models that performed the best in terms of the most recent machine translation competitions.

4.2.3 BERT-fused NMT

As an additional baseline, we decided to explore the technique proposed in (Zhu et al., 2020). The authors proposed a novel way of incorporating a large pretrained language model (BERT) into neural machine translation (NMT) using an extra attention model for both the NMT encoder and decoder. We used a transformer model architecture for this baseline. However, the results were not very promising. The results can be found in the Appendix.

4.3 IndoBART

(Cahyawijaya et al., 2021) built IndoBART, a pretrained encoder-decoder model based on the mBART model (Liu et al., 2020), but with different datasets and hyperparameter configurations. IndoBART is pretrained only on 'Javanese', 'Sundanese' and 'Indonesian' and lacks 'Minangkabau' training data.

We fine-tuned IndoBART for machine translation task on the parallel corpora for each respective language pairs. For the case of 'Minangkabau', this technique outperformed all others. We hypothesize that this might be due to the language similarities between 'Indonesian' and 'Minangkabau'. This phenomenon of similarity between the 'Indonesian' and 'Minangkabau' is because of high word overlap. Also, 'Indonesian' and 'Minangkabau' are mutually intelligible with some overlaps of lexicons and syntax, which is the main reason for their high similarity.

4.4 Multilingual Training

In multilingual training, we focused on combining the dictionaries of the languages we are using to actually take into account the commonalities of the languages to produce better set of results. To perform multilingual training we concatenated the parallel corpora of between the three languages 'Sundanese', 'Javanese', and 'Minangkabau' to 'In-

Language	Data	Base Model	5k sentences	10k sentences	15k sentences	20k sentences
jav-ind	val	-	19.87	20.96	21.48	**21.75**
	test	12.88	21.78	23.05	23.32	**23.83**
ind-jav	val	-	13.08	13.87	14.36	**14.62**
	test	7.30	12.28	12.22	**12.30**	12.23

Table 3: Ablation Study for the FLORES-101 dataset

Language	Data	20k sentences	5k backtrans jav	5k backtrans ind	5k backtrans jav + ind
jav-ind	val	21.75	21.98	**22.26**	22.03
	test	23.83	23.80	**23.90**	23.64
ind-jav	val	14.62	**17.21**	14.40	16.56
	test	12.23	**12.82**	11.88	11.76

Table 4: Backtranslation results with Malay as the pivot language on the Flores-101 dataset

donesian' and vice-versa. This did give a huge performance boost which is described in the last column of the Table 2. The model as we observe benefits from the addition of the 'Minangkabau' dictionary the most because of the high overlap in the dictionaries of 'Minangkabau' and 'Indonesian'.

The performance is significantly improved with multilingual training compared to the other approaches. We noticed higher improvements for translating to and from 'Sundanese' and 'Javanese' to 'Indonesian' than for translating to and from 'Minangkabau'. By concatenating the training data with similar language like 'Minangkabau', the model can learn more patterns from the closely related language because of the word overlap that exists between 'Indonesian' and 'Minangkabau'. However in case of 'Minangkabau' the language suffers from the increase in dictionary and thus performs worse as compared to just a bilingual task where the model only has to deal with the 'Minangkabau' dictionary.

5 Experimental Setup

5.1 Domain Knowledge Transfer

The dataset we use for testing the domain transferability of the model we had trained on the bible datset we perform an ablation study with the FLORES-101 dataset which consists of sentence pairs generated from wikipedia. Due to computational constraints, we are only using a subset of the FLORES-101 training dataset, particularly the jav-ind WikiMatrix dataset which contains 19,595 sentences (here after referenced as 20k). To give you a brief idea of the experimental setup we use the multilingual model based on the BART architecture, trained on the concatenated dataset of 'Javanese', 'Sundanese', 'Minangkabau', and 'Indonesian'. Using the model trained on the bible dataset we perform further fine-tuning on the FLORES-101 dataset, during this fine-tuning we slowly increase the number of sentences introduced to the training dataset for fine-tuning while keeping the test and val sets constant.

5.2 Data Augmentation

For this section we took inspiration from the session we had in class where we talked about using single pivot and double pivot languages to perform back translation on the base low resource language to augment the data with new parallel corpus pairs. Also during this exploration we looked at the paper (Edunov et al., 2018) mentioned for the assignment 2's page. We initially started off by looking at the various morphological similar languages to 'Javanese' like 'Malay' (msa) which is a cognate of 'Javanese' but has higher resource mapping to 'Indonesian' as compared to 'Javanese' itself. We use the google translate API to create backtranslation pipeline to try and improve the results generated for the FLORES-101 dataset. For generating more domain similar data we use 5k sentences from the FLORES-101 dataset and use the above mentioned backtranslation pipeline and we use 'Malay' as the pivot language in the following ways:

1. 'Javanese' -> 'Malay' -> 'Javanese' and map these newly generated sentences to the existing 'Indonesian' sentences to create new parallel data points for further discussion lets consider these sentence mappings as set 1. The

Reference	Base Model
Sejarawan pernah mengkritik kebijakan FBI di masa lalu, karena memfokuskan sumber daya pada kasus-kasus yang mudah diselesaikan khususnya kasus mobil curian, hanya untuk menaikkan tingkat keberhasilan agensi. FBI, keberhasilan (*success*), agensi (*agency*)	ahli silsilah yesus, yang pertama, dengan maksud untuk memberikan sumber daya kepada perkara yang gampang dig en apkan, terutama untuk kasus yang dic urinya, dengan tujuan untuk menambahkan keh inaan pada jalan kehidupan. yesus (*Jesus*), kehinaan (*humiliation*), jalan kehidupan (*path of life*)
Mereka semua lari dari tempat kecelakaan itu terjadi. kecelakaan (*accident*)	mereka lari dari tempat yang ked ur haka an itu. kedurhakaan (*iniquity*)
Pengurangannya telah dijadwalkan pada hari Selasa, tetapi diselamatkan setelah keputusan pengadilan darurat. hari Selasa (*Tuesday*)	ia telah dipilih untuk hari penghakiman, tetapi ia dipilih untuk selama keputusan pengadilan. hari penghakiman (*judgement day*)

Table 5: Appearance of biblical phrases in MT result of the non fine-tuned multilingual model

Reference	Base Model	Fine-Tuned (20k sentences)
Kelompok Studi Irak mempresentasikan laporannya pada pukul 12.00 GMT hari ini. Irak (*Iraq*)	penghuni - peng u asa di daerah itu menunjukkan lap arnya pada jam 12. daerah itu (*that place*)	sekelompok peneliti irak menunjukkan laporan pada pukul 12. 00 gmt.
Layaknya pasang surut yang terjadi akibat bulan menarik bumi, Bima Sakti memberi galaksi Sagitarius gaya. Bima Sakti (*Milky Way*) Sagitarius (*Sagittarius*)	seperti bulan memakai ruang atas bumi untuk pasang, dan juga bulan memakai kekuatan kepada bintang - b intang itu. bulan (*moon*) bintang-bintang (*stars*)	seperti misalnya, bulan sabit mendekati bumi, menyebabkan pasang, sem ana juga bima sakti berhubungan dengan galaksi sag itari us.
Setelah terjadi kecelakaan itu, Gibson dibawa ke rumah sakit tetapi tak lama kemudian meninggal.	selama peristiwa itu terjadi, ia telah diangkut ke rumah sakit, tetapi ia tidak lama mengalami hal itu. ia (*he*)	setelah tabrakan itu terjadi, gib son dibawa ke rumah sakit tetapi meninggal beberapa lama kemudian.

Table 6: Error Analysis for Named Entity Recognition in the MT result of FLORES-101 test dataset without fine-tuning and with fine-tuning with 20k sentences

following translation has a BLEU Score of 11.3089.

2. ‘Indonesian’ -> ‘Malay’ -> ‘Indonesian’ and map these newly generated sentences to the existing ‘Javanese’ sentences to create new parallel data points for further discussion lets consider these sentence mappings as set 2. The following translation has a BLEU Score of 38.2050.

Using these two sets we create an ablation study first by adding only set 1, then by only adding set 2, and then by augmenting the dataset by adding both the sets to the training data. The low BLEU score for ‘Javanese’ showcases the high relevance of our research, as even google translate has a tough time translating the mentioned low resource languages.

6 Analysis and Discussion

We discuss the performance of our models in multiple ways some of which are described below, we use the compare-mt tool presented in the paper (Neubig et al., 2019). This tool provides us with detailed comparison between two methods and also the accompanying BLEU scores.

Reference	Base Model	Fine-Tuned (20k sentences)
Lebih dari empat juta orang pergi ke Roma untuk menghadiri pemakaman. empat juta (*four million*)	lebih dari empat ribu manusia pergi ke roma. empat ribu (*four thousand*)	lebih dari empat juta manusia mara ke roma.
Beberapa layar televisi besar dipasang di berbagai tempat di Roma supaya orang-orang bisa menyaksikan upacara tersebut. beberapa (*several*)	dan sebuah layar lebar yang besar dipasang di beberapa tempat di roma, supaya mereka dapat menyaksikan apa yang telah diadakan - nya. sebuah (*a/one*)	beberapa layar tv dipasang di beberapa lokasi di roma untuk melihat upacara tersebut.
Mereka termasuk Belanda, dengan Anna Jochemsen selesai di urutan kesembilan di kelas pengelompokan putri di Super-G kemarin, dan Finlandia dengan Katja Saarinen selesai di urutan kesepuluh di acara yang sama. kesembilan (*ninth*), kesepuluh (*tenth*)	mereka termasuk seorang p elah ap dengan anna j ara yang keempat di dalam kelas yang sama, dan di tempat yang sama dengan dia j ara untuk ketiga kalinya di dalam pertandingan yang sama. ketiga (*third*), keempat (*fourth*)	ia adalah belanda, dengan anna jo chem sen finis ke - 10 di kelas mendirikan wanita di super - g w ingi, dan finlandia dengan ket atan egaraan sepuluh di dalam acara yang sama. ke-10 (*10th*), sepuluh (*ten*)
Itu adalah pertandingan terakhir bagi All Blacks, yang sudah memenangkan trofi tersebut dua pekan lalu. dua pekan lalu (*two weeks ago*)	inilah yang terakhir bagi dia, yang telah menang dua minggu lamanya. dua minggu lamanya (*for two weeks*)	ini adalah pert andh ingan terakhir untuk all black s, yang akan memenangi piala dua minggu sekali. dua minggu sekali (*once every two weeks*)

Table 7: Numerical errors in the MT result of FLORES-101 test dataset without fine-tuning and with fine-tuning with 20k sentences

6.1 Multilingual training Impact

We see a noticeable impact of multilingual training on our models performance, especially in case of languages like ‘Javanese’ and ‘Sundanese’. The reasoning behind why we see such a phenomenon is because of the similarity between ‘Sundanese’ and ‘Javanese’ the training data now contains more diverse examples for similar words. In case of ‘Minangkabau’ we notice the opposite the BLEU score reduces as compared to improving. Our thinking as to why this might be happening is because ‘Indonesian’ and ‘Minangkabau’ are mutually intelligible with some overlaps of lexicons and syntax, resulting in the initial high min-ind translation score. The decrease we feel is caused by the addition of new data points in the dictionary from ‘Sundanese’ and ‘Javanese’ which could be causing confusion for the model.

We also researched the lexicostatistics analysis for these languages to back up our observations.

Language	jav	sun	min	msa
jav	-	-	-	-
sun	49	-	-	-
min	37	34	-	-
msa	45	38	61	-

Table 8: Lexicostatistics results from the papers (Hafizah, 2018) and (Suyata, 2012)

Lexicostatistics is a method of comparative linguistics that involves comparing the percentage of lexical cognates between languages to determine their relationship as can be seen in Table 8. For the usecase of our scenario we researched the following papers, (Hafizah, 2018) and (Suyata, 2012). Since ‘Indonesian’ is a standardized version of ‘Malay’, the comparison of between these languages with ‘Malay’ and ‘Indonesian’ should be similar. The lexicostatistics between ‘Malay’ and ‘Minangkabau’ have the highest score at 61 compared to 'Javanese' and 'Sundanese' showing that these two languages have more overlap than the others. 'Malay' in this use case acts as a substitute of 'Indonesian' because of the unavailability of lexicostatistic results for 'Indonesian' with the other languages. These results and the Table 8 thus back our findings as to why we see such high results for 'Minangkabau' to 'Indonesian' machine translation as compared to the other two languages.

6.2 Domain Knowledge Transfer

As can be seen on Table 3, using the multilingual model trained on bible data with FLORES-101 test dataset results in poor performance due to small training data size and domain mismatch. The multilingual model was trained on the Bible dataset which is smaller and domain-specific compared to the FLORES-101 dataset. As we add more training data from the same domain, we can see an increase in BLEU score. This domain mismatch problem can be seen by a number of biblical phrases showing up in the translation result (Table 5). One of the errors that we observed is Named Entity Recognition (NER) error as detailed in Table 6. In the results of the non fine-tuned model, named entity such as a name of a person is substituted with a pronoun. With the fine-tuned model, named entities are not substituted despite having some tokenization error. Another error we have observed are numerical errors as seen in Table 7. Despite both models producing sentences with numerical errors, the error rate is higher for the non fine-tuned model. One reasoning for the high numerical error in the non fine-tuned model is possibly the way the numbers are represented in the Bible as compared to how we normally use them words like 'million' don't appear with much frequency in the Bible, most times in the Bible words like 'hundred thousand' is used. Another such example is the use of phrases like 'two of every' in the Bible which can lead to the errors we see in our Wikipedia dataset, as no one uses phrases like these in todays world.

6.3 Data Augmentation

Using the multilingual model described in Section 4.4, we fine-tuned the model with FLORES-101 20k parallel data concatenated with the backtranslated results described in Section 5.2. As seen on Table 4, with the addition of backtranslated data, the translation performance only gain a slight increase. This might be caused by google translate result not being good enough for these languages. Using backtranslated data in addition to the 20k training sentences slightly improved the performance for the respective target language, adding 'Indonesian' backtraslated sentences improved jav-ind performance while adding 'Javanese' sentences improved ind-jav results. The addition of backtranslated sentences in target language added more variety of sentences and expanded the vocabulary. Adding backtranslated source language did not improve or even deteriorate the translation result.

For jav-ind translation, using backtranslated 'Indonesian' gave slightly better performance than using backtranslated 'Javanese'. The main reason for that is the low BLEU scores for the backtranslated 'Javanese' which is 11.3089 as compared to that of 'Indonesian' is 38.2050. For most sentences the translation result is similar, having errors such as word mismatch at the exact same location. These errors might be due to out of vocabulary words that does not exist in the training data. For some sentences, the translation result using backtranslated Javanese produce more error where they contain more 'Javanese' words as can be seen in Table 9.

7 Conclusion

Throughout this project our main focus was on improving the state of neural machine translation for the low-resource Indonesian languages, especially the ones in the Western Malayo-Polynesian region. The reasoning behind choosing these languages was two fold, one they don't have a lot of work done on them and second they are similar to each other, thus we leverage this similarity to progress the field overall. The results produced in this report showcase a couple of intuitions we had at the start of this project, we had assumed that performing multilingual training using similar languages with high dictionary would be essential. Using a pre-trained language model to build upon also proved to be important for our machine translation task. Backtranslation however underperformed our expectations, one of the reasons we feel for that is because the existing google translate isn't good enough which makes our work even more important. The main takeaway from this project is that using multilingual machine translation to improve the current set of resources available for low resource languages. Looking towards the future we want to actually build upon our current set of results and evaluate our models with a different metric other than the BLEU score. We would also like

Reference	backtrans ind	backtrans jav
Akibatnya, para pemainnya mengisap ganja di atas panggung, dan teater itu sendiri mendorong penontonnya untuk ikut bergabung.	sebagai balasannya, seniman mengusung nya ganja di panggung, dan teater tersebut mengajak penonton untuk turut serta menghasut.	as ile, seorang seniman mengusung teater di panggung, dan teater ini mengundang para penonton untuk ikut bergabung.
Pada dasarnya dia (Wales) telah berbohong kepada kami sejak awal. Pertama, dengan bertindak seolah-olah ini untuk alasan hukum. Kedua, dengan berpura-pura dia mendengarkan kami, hingga penghapusan karya seninya.	dh ewek e (lan ang) [wal es] dasar sangat ngap usi terhadri awal. pertama, dengan melakukan seperti - k aya ini untuk alasan hukum. kedua, dengan berturut - t urut (dan ang) meminta kita, hingga dengan penghapusan sen inya.	d ewa e (lan ang) [wal es] dasar mereka mengancam kita dari awal. sep isan, dengan tumindak seperti - k aya ini untuk alasan hukum. kaping pind ho, dengan et hok - eth ok dia (dan ang) ngr ung o ake kita, sampai dengan penghapusan sen ine.

Table 9: Translation result differences between data augmentation with backtranslated Indonesian and backtranslated Javanese on jav-ind translation task. Indonesian words are denoted in blue and Javanese words are denoted in red.

to explore other techniques which could be helpful in machine translation for low resource languages like regularized fine-tuning techniques.

References

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: Nlp challenges for underrepresented languages and dialects in indonesia.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Leylia Khodra, Ayu Purwarianti, and Pascale Fung. 2021. Indonlg: Benchmark and resources for evaluating indonesian natural language generation.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Hafizah Hafizah. 2018. Leksikostatistik bahasa indonesia dengan bahasa minang dialek bukittinggi (kajian linguistik historis komparatif). *DEIKSIS*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *CoRR*, abs/1701.02810.

Fajri Koto and Ikhwan Koto. 2020. Towards computational linguistics in Minangkabau language: Studies on sentiment analysis and machine translation. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 138–148, Hanoi, Vietnam. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *CoRR*, abs/2001.08210.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, Xinyi Wang, and John Wieting. 2019. compare-mt: A tool for holistic comparison of language generation systems. *CoRR*, abs/1903.07926.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *CoRR*, abs/1904.01038.

Pujiati Suyata. 2012. Dari leksikostatistik ke glotokronologi : Analisis sembilan bahasa di indonesia. *Humaniora*, 11(1):69–75.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating BERT into neural machine translation. *CoRR*, abs/2002.06823.

A Appendix

Table 10 contains results for the BERT-fused NMT approach. Due to the computational resource and

Language	**BLEU**
ind-jav	17.43
min-ind	63.55
ind-min	52.36

Table 10: BLEU scores for BERT-fused NMT

time constraints, these experiments were not conducted for all language pairs. We hypothesize that with more training epochs, this approach would at best give comparable results to the baselines and thus, we decided to limit this model to 3 language pairs.

B Work Distribution

Sidharth worked on the OpenNMT baseline and backtranslation pipeline. Aditi worked on the Fairseq baseline and the BERT-fused method. Venny worked on IndoBART fine-tuning and multilingual training. Data preprocessing and FLORES ablation are divided equally. Since Venny is a native speaker of Indonesian, she did most of the error analysis involving looking at the translation results. All members worked on other parts of the report including analysis.