Analysis of Language Model Role in Improving Machine Translation Accuracy for Extremely Low Resource Languages

Herry Sujaini [1,*], Samuel Cahyawijaya [2], and Arif B. Putra [1]

[1] Department of Informatics, University of Tanjungpura, Pontianak, Indonesia; Email: arifbpn@untan.ac.id (A.B.P.)
[2] Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Kowloon, Hong Kong; Email: scahyawijaya@ust.hk (S.C.)
*Correspondence: hs@untan.ac.id (H.S.)

***Abstract*—Several previous studies have suggested using statistical machine translation instead of neural machine translation for extremely low-resource languages. We could translate texts from 12 different regional languages into Indonesian using machine translation experiments. We increased the accuracy of machine translation for 12 extremely low-resource languages by using several monolingual corpus sizes on the language model's target side. Since many Indonesian sources are available, we added this corpus to improve the model's performance. Our study aims to analyze and evaluate the impact of different language models trained on various monolingual corpus on the accuracy of machine translation. The increase in accuracy when enlarging the monolingual corpus is not observed every time, according to our experiments. Therefore, it is necessary to perform several experiments to determine the monolingual corpus to optimize the quality. Experiments showed that Melayu Pontianak achieved the highest bilingual evaluation understudy improvement point. Specifically, we found that by adding a monolingual corpus of 50–100K, they performed a bilingual evaluation understudy improvement point of 2.15, the highest improvement point they reached for any of the twelve languages tested.**

***Keywords*—statistical machine translation, extremely low resource languages, monolingual corpus, language model, Indonesian**

I. Introduction

Machine translation is the automated process of translating text or speech from one language to another using computational techniques. Machine translation holds significant potential in Indonesia, a linguistically diverse country with numerous regional languages. Indonesia boasts a rich tapestry of regional languages with distinct cultural heritage and linguistic nuances. Applying machine translation technology to Indonesia's regional languages can bridge communication gaps, fostering greater understanding and connectivity among its diverse population. By developing sophisticated machine translation systems that accurately capture the subtleties of these regional languages, Indonesia can promote inclusivity, facilitate trade and tourism, and enable knowledge sharing across its vast linguistic landscape. Through the power of machine translation, Indonesia can embrace its linguistic diversity and unlock a world of opportunities for its people.

A. Machine Translation

Creating Machine Translation (MT) systems aim to automate text translation from one language to another. MT is done to improve communication between people who speak different languages and facilitate access to information hampered by language differences. Machine translation has practical applications in various fields, including international business, tourism, and cross-cultural communication. It can also help preserve endangered languages by enabling language custodians to translate and share important cultural documents and stories. Machine translation can also help language education by giving students instant translation and allowing them to practice reading and writing in a foreign language.

Here are some issues to deal with when building a machine translation system, including:

(1) Lack of Data: One of the main hurdles is finding enough high-quality, parallel data to train machine learning models. In many languages, especially low-resource languages, annotated data is limited, and it is challenging to prepare an accurate model;
(2) Language Ambiguity: Natural languages are inherently ambiguous, making it difficult to determine the correct translation in some situations. For example, the same word can have different meanings in different contexts, or words can have several meanings in the same context;
(3) Cultural Differences: Cultural differences between languages can also make machine translation difficult. Some phrases, jokes, and sayings may not have an equivalent in the target language, making it difficult to translate them accurately;
(4) Domain specificity: Another barrier is domain specificity, where the language used in a particular

Manuscript received March 5, 2023; revised April 11, 2023; accepted June 25, 2023; published October 13, 2023.

domain (such as law, medicine, or finance) may be highly technical and specialist. It requires special knowledge and expertise to translate accurately; and

(5) Model Complexity: Building a machine Translation Model (TM) can be complex and requires sophisticated algorithms to handle large data and computational resources.

Most translation machines already use Neural Machine Translation (NMT) because it performs better. Nonetheless, several studies report that Statistical Machine Translation (SMT) is still better than NMT in language translation with low resources. They recommended using SMT instead of NMT for Extremely Low-Resource Languages (ELRL). These languages have minimal training data or resources, making developing machine translation systems that produce high-accuracy translations challenging. It has been recommended because SMT relies on statistical models that can efficiently handle small data sets and produce reliable translations for ELRLs. In contrast, NMT models require significant training data to achieve optimal performance and may be less suitable for ELRLs with limited resources. SMT is a promising approach for the machine translation of ELRLs [1–4].

In SMT, a Language Model (LM) plays a crucial role in predicting the probability of a sequence of words in a target language. The LM aims to ensure the translated sentence is grammatically accurate, meaningful and matches the target language's writing style and conventions. The LM uses statistical methods to estimate the probability of a sequence of words in the target language. It does this by considering the frequency of word combinations in a large training corpus. The model then uses this information to assign a probability to each possible word in the target language at each time step. This information is combined with the TM's outputs to generate the final translation. The LM helps to address the issues of under and over-generation that can occur in SMT, where the TM may generate sentences that are too short or too long. The LM ensures the translated sentence is correct and coherent, maintaining fluency and meaning by checking syntax and semantics.

B. Indonesia's Regional Languages

Indonesia is a diverse country with over 270 million people living across 18,000+ islands and speaking many different languages. Indonesia is a country that has the second largest variety of regional languages in the world after Papua New Guinea, namely 726 languages, covering 10% of the world's languages, each with its unique grammatical characteristics. However, some common features of these languages include using affixes to indicate tense, aspect, and mood and the absence of grammatical gender. Many regional languages also have a simple sentence structure and a flexible word order. Additionally, some regional languages in Indonesia use reduplication, or the repetition of a word or a part of a word, to express various grammatical meanings [5–8].

Indonesia's regional languages exhibit various phonetic and phonological characteristics, including tone, vowel, consonant systems, and stress patterns. These languages have incorporated words from neighboring languages, Dutch, Arabic, and other sources. Syntactic structures of regional languages in Indonesia differ, including agreement systems, case marking, and word order. Several regional languages have historically been documented using native scripts such as Jawa or Sunda. Others use Latin-based alphabets. Many languages in Indonesia are endangered because the national language of Indonesian has had a significant impact, and the younger generation is not practicing traditional languages.

C. Paper's Objectives

From the explanation above, it is necessary to analyze the right ways and approaches to implement machine translation in ELRL. In this study, we used SMT and implemented the strategy of increasing the MC size in the target language. The research has contributed new knowledge to machine translation by exploring the best approach to using monolingual corpus data for low-resource languages. By analyzing the impact of different monolingual corpus sizes on translation accuracy for various languages, the study provides insights into the optimal approach for machine translation. Additionally, the research sheds light on each language's specific challenges by comparing the accuracy of translations for different languages using the same amount of monolingual corpus data. It highlights areas for improvement in machine translation technology.

The limitations of this study involve obtaining sufficient high-quality data for languages with few resources, dealing with the natural ambiguity of languages that complicates finding the correct translations, and the difficulty of translating technical language in specialized areas that need expert knowledge. Additionally, the study's focus on 12 extremely low-resource languages limits the scope of its findings. Exploring the impact of different MC sizes on translation accuracy may not cover all factors influencing translation quality.

The contribution of this paper lies in its exploration of improving TM accuracy for 12 extremely low-resource languages by incorporating various MC sizes on the target side of the language model. The study provides insights into the impact of different LM, trained on varying MC, on machine translation accuracy by conducting multiple experiments. The results show that increasing the size of the MC does not constantly improve accuracy, indicating the need for further experimentation to determine the optimal corpus size for each language.

II. Literature Review

To construct machine translation systems effectively, a mix of linguistic knowledge, state-of-the-art machine learning techniques, and extensive amounts of comparable premium data are necessary [9]. The most significant challenge is finding the right balance between precision, velocity, and expense to create a system that can translate text accurately and practically [10, 11]. The biggest challenge for implementing machine translation is

the data set used to build the model. Researchers use several approaches to build machine translation systems, including:

(1) Rule-Based Machine Translation (RBMT): This method uses pre-established grammar and vocabulary regulations to translate text from one language to another. RBTM is a simple strategy that can provide high accuracy, but it requires a lot of time to create and keep up, and it is unable to process novel or unfamiliar words or expressions [12];
(2) SMT: This technique employs statistical models to translate text based on patterns found in an extensive parallel text corpus. The SMT models learn to translate words and expressions based on their surroundings and can manage novel or unfamiliar words and phrases. SMT is commonly utilized in commercial machine translation systems and is renowned for its speed and precision [13, 14];
(3) NMT: This method utilizes deep neural networks to translate text. NMT models are trained using extensive parallel text corpus and can handle the intricacies of human languages, such as grammar, meaning, and context. NMT has become the leading-edge technique in machine translation and is recognized for its exceptional accuracy and fluency [12, 14].

These are among the primary methods used to build machine translation systems, but not the only ones. Depending on the application's needs, systems may combine these methods or use only one. Factors such as the amount of data, target languages, required accuracy, and available computational resources influence the method chosen. A high-quality Parallel Corpus (PC) with large data is necessary to develop a translation engine for each language pair. Building a language translation machine with minimal resources is challenging, as collecting the necessary data requires substantial cost and time. Certain researchers use an alternative method to create translation machines with limited or no resources. This technique entails using an intermediary language, but it necessitates having access to parallel datasets that include translations from the source language to the intermediary language and from the intermediary language to the target language [15]. Languages without available resources cannot use this approach for translation into other languages. Another approach used for low-resource languages is like back-translation [16, 17], multilingual knowledge transfer [17, 18], and unsupervised NMT [19], depending on either parallel corpus of different languages or a significant amount of monolingual data for the language of interest [20].

In general, NMT models require large amounts of parallel data, i.e., data with corresponding sentences in both the source and target languages, to train effectively. The quality of the NMT system may be compromised if limited data is available for a low-resource language. However, some techniques, such as data augmentation, transfer learning, or multi-task learning, can overcome this limitation. These techniques allow NMT models to be trained on a small data set and still achieve good results by leveraging information from related languages or other tasks. Additionally, pre-training large NMT models on a large corpus of text data in multiple languages and then fine-tuning them on a smaller data set for a low-resource language can also lead to good results. In summary, NMT is a powerful approach to machine translation. Still, its performance for low-resource languages will depend on data availability, computational resources, and the specific techniques used to build the system [21].

SMT is a machine translation technique that employs statistical models to convert text from one language to another. It functions by acquiring knowledge of the association between words and phrases in the source and target languages, using a massive parallel corpus of text. The resulting models are then used to translate new text based on the patterns learned from the data. SMT is a widely used approach for machine translation, and it has successfully produced high-quality translations for many language pairs. It is known for its speed, scalability, and ability to handle rare or unknown words and phrases. SMT is a flexible approach that can be adapted to different languages and domains. It has been used in various applications, including commercial machine translation systems, information retrieval, and multilingual information access.

Advantages of SMT over NMT include (1) Simplicity: SMT is a simpler approach than NMT, which makes it easier to understand and debug; (2) Speed: SMT models are faster to train and faster to translate than NMT models, which makes them more suitable for real-time translation applications; and (3) Handling of rare or unknown words: SMT models are better equipped to handle rare or unknown words, as they can translate based on word-level information, such as part-of-speech or morphological information.

An effective method for improving the quality of SMT translations with very low resources is to utilize the target language's Language Model (LM) [22]. Several studies have proven that LM has a role in improving translation results, for example, phrase-based unsupervised and semi-supervised MT [19, 23–26]. Increasing the quantity of the MC, especially Indonesian, is not too difficult because many resources are available.

Low-resource language has limited amounts of annotated data, computational resources, or linguistic information. The above can challenge building machine learning models for natural language processing tasks in these languages. Many researchers have defined Low Resource Language (LRL) in Natural Language Processing (NLP), including the problem of low resources that can arise mainly due to languages that considered low resources or domains that are considered low resources [27]. LRL is often called an under-resourced, low-density, resource-poor, low-data, or under-resourced language. These terms appear depending on the viewpoint of different scenarios and resource conditions [28]. Ghafoor *et al.* [29] examine the language

assets of social media users who have friends and followers primarily from the same community or country and prefer to communicate in their native language to exchange viewpoints. English is the most widely used language on the internet, followed by Chinese and Spanish. However, users also communicate in other languages, such as Arabic, Indonesian, Malaysian, Portuguese, French, Hindi, Urdu, and more, which are referred to as Low-Resource Languages (LRLs).

A PC in machine translation is a collection of texts in two or more languages with corresponding sentences or passages in each language. These texts are used to train and evaluate machine translation systems. The PC serves as the input for the machine learning algorithms that generate the translations, and the output's quality depends on the PC's size, quality, and relevance. The PC provides the model with examples of the source language sentence and its corresponding target sentence, allowing it to learn patterns and relationships between the two languages. Research in the field of MT and NLP is more dominant in English and several languages on the European continent because resources can be more easily obtained. This is inversely proportional to regional languages around the world, which have very few resources. These languages are called ELRL. Regarding quantity, some mention them with a limit of 0–13 K sentences [30].

Machine translation has been crucial in facilitating communication across cultural and national boundaries in recent years. As the demand for machine translation systems grows, the challenge of producing accurate and optimal language-process translations remains. Although various machine translation systems exist, the quality of translations still requires improvement. This literature review examines research on machine translation involving Indonesian and other languages, highlighting different approaches, tools, and evaluation methods used to measure performance. Furthermore, it proposes future work to enhance machine translation quality between Indonesian and other languages. The review findings reveal that attention-based approaches are increasingly employed to improve neural machine translation performance. The quality of translation is influenced by factors such as corpus size, well-aligned corpora, and the techniques applied [31].

There exist gaps in machine translation techniques for low-resource languages due to limited data, computational resources, or linguistic information. These challenges make building effective machine learning models for natural language processing tasks in such languages difficult. Alternative methods for low-resource languages, such as intermediary languages, back-translation, multilingual knowledge transfer, and unsupervised NMT, rely on parallel corpus or substantial monolingual data, which may be scarce for these languages.

Possible solutions to address these gaps include data augmentation, transfer learning, or multi-task learning. These techniques enable NMT models to train on small datasets while leveraging information from related languages or tasks to achieve better results. Pre-training large NMT models on a vast corpus of text data in multiple languages and fine-tuning them on smaller datasets for low-resource languages can also yield good results. SMT is another approach that can be used to translate low-resource languages. It is simpler, faster, and better equipped to handle rare or unknown words compared to NMT. Utilizing the target language's LM has been proven effective in improving translation quality for low-resource languages. Increasing the monolingual corpus, especially for languages with many available resources, can further enhance the performance of SMT models.

In summary, addressing the gaps in machine translation for low-resource languages requires exploring alternative techniques, leveraging information from related languages or tasks, and utilizing the target language's language model. These strategies can help improve translation quality and make machine translation more accessible for a broader range of languages.

III. Materials and Methods

We conducted experiments using MOSES [13]. MOSES is a statistical-based machine translation framework widely used by researchers and a robust and widely used framework for building machine translation systems using SMT techniques. SMT requires two types of data sets: parallel corpus and monolingual corpus. The parallel corpus generates the translator model, while the monolingual corpus creates the LM on the target side.

A. Materials

We have used the PC sourced from the Indonesian NLP Data Catalogue. This catalog is provided by Nusa crowd [32], a collaborative initiative to collect and unite existing resources for Indonesian languages. There are 24 parallel corpus of regional languages available from Korpus Nusantara [22], a vast collection of written and spoken texts in various Indonesian languages. For our study, we chose to work with ten of these languages, namely: Javanese Kromo (JK), Javanese Ngoko (JN), Dayak Ahe (DA), Dayak Taman (DT), Batak Toba (BT), Melayu Ketapang (MK), Melayu Pontianak (MP), Melayu Sambas (MS), Madura (MA) dan Tiociu Pontianak (TP). Apart from these ten languages, we also use the Sundanese parallel corpus (SU) from the "Sundanese-Indonesian PC" [33], dan Bahasa Minang (MI) from "MinangNLP MT" [34]. These languages were selected based on several factors, such as their prevalence in the region, the availability of parallel corpus, and their similarity to the Indonesian language. We aimed to include a diverse range of languages, with varying levels of resources available, to evaluate the performance of our machine translation system across different scenarios.

By working with these ten languages, we obtained a representative sample of the languages spoken in the region, which would help us draw meaningful conclusions about the effectiveness of our approach. Overall, our study aimed to contribute to machine translation, particularly in the context of low-resource

languages, by providing insights into the performance of SMT models when combined with monolingual corpus.

In addition to the PC, we utilized an Indonesian MC from the Leipzig Corpus Collection [35], which contained a dataset of one million Indonesian sentences. We used this corpus to generate the LM on the target side in our machine translation system. By incorporating this large dataset of monolingual Indonesian text, we aimed to improve the accuracy of the translations, particularly in cases where the PC was limited. The Leipzig Corpus Collection is a well-known and widely-used resource in natural language processing and is known for its high-quality and diverse range of texts. By utilizing this corpus, we could use the vast amount of available data to train our machine TM, allowing us to produce more accurate and reliable translations.

Table I shows the characteristics of each PC we used in our study. We investigated 12 regional languages for this research, and the total number of sentences we used was 67K. The number of tokens in each sentence varied between 3 and 25.

Fig. 1 presents a comparison of token counts and vocabulary sizes between Indonesian and local languages. One intriguing observation from this data is that nearly all local languages possess a higher number of tokens than Indonesian, except for the MI and MA languages. This suggests that regional languages in Indonesia are generally more concise than the Indonesian language. Similarly, regarding vocabulary size, only the MI and MA languages have larger vocabularies than Indonesian. This implies that word variation in regional languages is not as diverse as it is in Indonesian. One reason for this could be that the Indonesian language incorporates many words from other languages, including those from regional languages.

TABLE I. CHARACTERISTICS OF REGIONAL LANGUAGE-INDONESIAN PARALLEL CORPUS

No	ID	Languages	Sentences	Indonesian		Regional	
				Tokens	Vocabularies	Tokens	Vocabularies
1	MI	Minang	16,371	204,932	21,258	204,468	25,204
2	MS	Melayu Sambas	9099	66,194	8700	68,000	11,305
3	BT	Batak Toba	6909	66,469	6695	67,211	7861
4	JN	Jawa Ngoko	6059	60,075	6018	60,568	7659
5	MK	Melayu Ketapang	5189	62,890	9040	63,194	10,033
6	JK	Jawa Kromo	5000	44,804	7080	51,691	14,172
7	TP	Tiociu Pontianak	5000	22,300	3269	24,450	1467
8	MP	Melayu Pontianak	3746	30,602	4920	30,636	5909
9	SU	Sunda	3616	43,889	7125	48,662	8234
10	DT	Dayak Taman	3109	18,695	1812	19,726	3178
11	DA	Dayak Ahe	1994	26,466	3339	28,678	4505
12	MA	Madura	1100	14,493	2236	14,466	2395

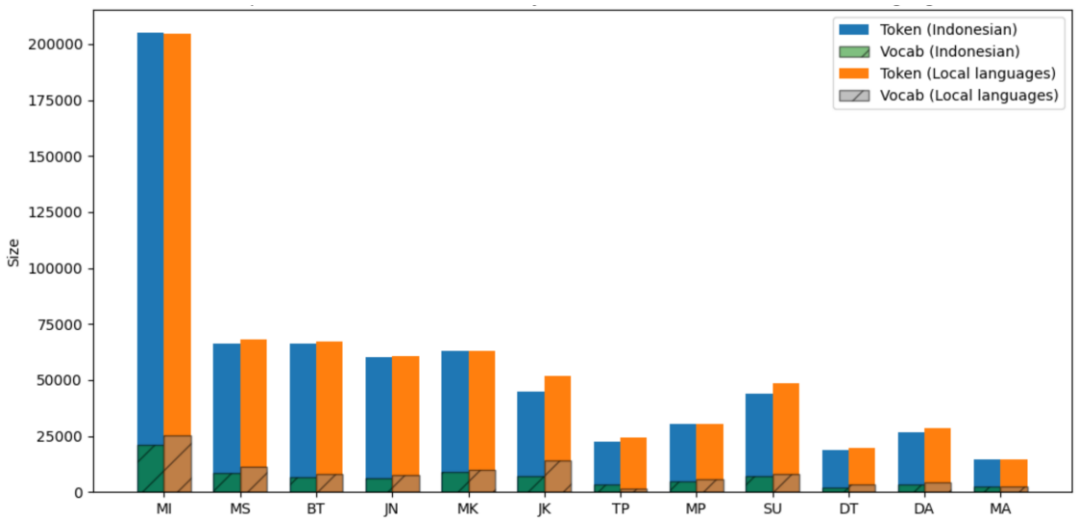

Figure 1. Comparison of tokens and vocabulary size between Indonesian and local languages.

B. Methods

In SMT, pre-processing is an important step that prepares text data for further analysis and modeling. The primary purpose of pre-processing is to ensure that text data is consistent, structured, and easier to process. We carry out three pre-processing: cleaning, tokenization, and case folding.

1. Cleaning: This step involves removing unwanted characters, symbols, or elements from text data that could hinder translation. This may include HTML tags, URLs, special characters, or other irrelevant information. In addition, we also check whether the number of sentences in the source language is the same as the target language.
2. Tokenization: Tokenization is breaking text into individual words or tokens. This step is essential

because it allows the machine translation model to process and analyze the text at the word level.

3. Case folding: Converting all text data to lowercase is a typical pre-processing step in many natural languages processing tasks, including machine translation. This step ensures that the model treats words with the same meaning but different capitalization as the same sign. For example, "Paper" and "paper" would be considered distinct tokens without lowercase letters, but they would be treated as the same token after the lowercase letters. This step helps reduce data dimensions and simplifies the translation process.

Our experiment involved building 12 SMT models, one for each regional language, with Indonesian as the target language. To evaluate the impact of the LM, we built an additional 15 models with an increased quantity of language data. We gradually added more sentences to the MC for each model to observe the effect on the translation accuracy.

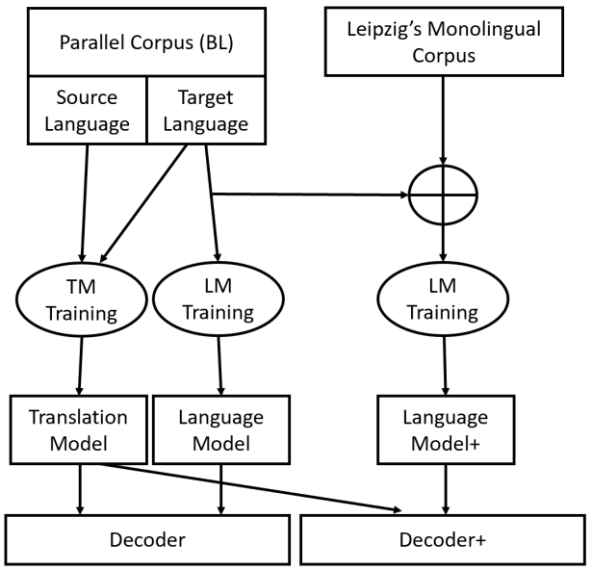

Figure 2. The architecture of strategy research on SMT.

The architecture of the machine translation model we implemented and the research strategy presented in Fig. 2 outlines the overall approach taken in this research. We use target language data from a PC to train an LM as a baseline. We train the source and target languages, prepared in parallel in the corpus, to produce TMs. The LM trained with SRILM and TM with Giza++ is then used in the decoder to translate sentences from regional languages to Indonesian. The resulting TM from the training is still used for machines with the same language but with a different LM (LM+). We obtained this LM+ by training the MCs by combining PC (Indonesian) sentences with additional Leipzig MCs. We use LM+ and TM from the previous training for decoder+ to translate sentences from regional languages into Indonesian using the same test data.

We did this for every 12 local languages that were translated into Indonesian. Then, we tested it again using the same test data with the increased quantity of LM+. We doubled the size of the MC to build the language model in increments of 10K starting from 1K, 10K, 20K, 30K, 40K, and 50K. We then continued increasing it by 50K to reach 100K, 150K, 200K, and 250K and finally doubled it to reach 500K and 1M sentences.

When choosing settings, we relied on prior knowledge from earlier research, focusing on settings that have proven successful in similar tasks or problems. We use several hyperparameters in this study, namely:

- UnknownWordPenalty0 = 1
- WordPenalty0 = −1
- PhrasePenalty0 = 0.2
- TranslationModel0 = 0.2 0.2 0.2 0.2
- Distortion0 = 0.3
- LM0 = 0.1

This configuration specifies a machine translation model's dense weights for various feature functions. The weights are used to score candidate translations during the decoding process.

- UnknownWordPenalty0: This is the penalty applied to unknown words in the translation. A value of 1 means a higher penalty is applied to translations with more unknown words.
- WordPenalty0: This is the penalty applied per word in the translation. A negative value (−1) indicates that shorter translations are preferred.
- PhrasePenalty0: This is the penalty applied to each phrase in the translation. A value of 0.2 encourages the use of fewer phrases in the translation.
- TranslationModel0: These are the weights for different features of the translation model. The four values (0.2, 0.2, 0.2, 0.2) indicate that each feature contributes equally to the overall score.
- Distortion0: This is the weight for the distortion model, which controls the reordering of words or phrases during translation. A value of 0.3 means that the distortion model moderately impacts the translation score.
- LM0: This is the weight for the Language Model (LM), which measures the fluency of the generated translation. A value of 0.1 indicates that the LM has a relatively low impact on the overall translation score.

IV. Result and Discussion

Bilingual Evaluation Understudy (BLEU) is a commonly used evaluation metric for machine translation that gauges the quality of machine-generated translations by comparing them to a set of reference translations. The BLEU point ranges from 0 to 1, with a score of 1 indicating a perfect match between the machine-generated translation and the reference translations. The BLEU point is based on precision, measuring the degree of word-to-word match between machine-generated and reference translations. The score also accounts for n-grams (short sequences of words) present in the reference translations, with greater weight given to longer n-grams. The BLEU helps to measure the fluency and coherence of the machine-generated translations. Although the BLEU point is a standard evaluation metric and widely used in machine translation research, it has some drawbacks,

such as not considering the semantic meaning of words and the context of the translations.

Table II illustrates the increase in MC data from 1K to 1M, which can be seen more clearly in Fig. 3. The illustration shows that every language in the data undergoes a consistent increase. Even with adding 50K sentences in the MC, the difference in the BLEU score compared to the baseline also increased by an average of 0.009. The MP language exhibited the highest increase, with a BLEU score difference of 0.022, while the JK language demonstrated the lowest increase of 0.004. Research on Indonesian-Lampung within the range below 50K showed an increase in BLEU score of 0.043 from 1K to 3K sentences in the MC. [24]. Research on Javanese-Indonesian shows an increase in BLEU points of 0.005 when increasing from 1K to 5K sentences in the MC [25]. Similarly, research on Muna-Indonesia shows an increase in BLEU points of 0.014 when increasing from 342 to 1351 sentences in the MC [26].

In the range of adding 50K to 250K sentences to the MC, it can be observed that not all languages experienced an increase in BLEU points, and some even decreased, such as MI, MK, JK, and DT. On average, the difference between the addition of 250K to 50K was 0.003. The highest score was obtained for BT, with a score difference of 0.015. The lowest score for the DT language was obtained, with a score of −0.008. After increasing the MC size again to 1M sentences, the average difference between adding 1M to 250K sentences in the MC was 0.002, with the highest score obtained by the DT language with a difference score of 0.010 and the lowest score obtained by the MK language with a score of −0.01 An Interesting thing happened with the DT language. Namely, the addition of 250K experienced the highest decrease, while increasing it to 1M resulted in the highest increase. The data analysis shows that although adding MC can generally improve the accuracy of SMT, it still varies with the dataset of each language.

TABLE II. BLEU POINTS 12 LANGUAGES WITH LM ENHANCEMENT

ID	BL	1K	10K	20K	30K	40K	50K	100K	150K	200K	250K	500K	1M
MI	0.743	0.747	0.748	0.748	0.750	0.750	0.751	0.751	0.751	0.751	0.751	0.752	0.752
MS	0.294	0.295	0.301	0.302	0.304	0.304	0.306	0.306	0.308	0.308	0.308	0.308	0.308
BT	0.805	0.808	0.813	0.811	0.812	0.810	0.812	0.818	0.821	0.824	0.826	0.831	0.830
JN	0.446	0.448	0.455	0.457	0.459	0.457	0.457	0.460	0.463	0.462	0.466	0.469	0.466
MK	0.622	0.626	0.634	0.635	0.636	0.637	0.636	0.631	0.628	0.627	0.629	0.627	0.619
JK	0.082	0.083	0.086	0.086	0.086	0.086	0.086	0.082	0.084	0.083	0.085	0.083	0.082
TP	0.192	0.188	0.190	0.188	0.193	0.195	0.196	0.196	0.198	0.200	0.202	0.209	0.217
MP	0.598	0.603	0.612	0.616	0.619	0.619	0.620	0.620	0.623	0.624	0.625	0.625	0.625
SU	0.536	0.540	0.551	0.550	0.551	0.551	0.547	0.552	0.552	0.552	0.549	0.551	0.553
DT	0.435	0.438	0.438	0.442	0.444	0.443	0.444	0.443	0.445	0.439	0.437	0.444	0.446
DA	0.423	0.423	0.427	0.427	0.425	0.429	0.428	0.427	0.437	0.437	0.437	0.438	0.443
MA	0.767	0.768	0.770	0.772	0.771	0.771	0.772	0.771	0.773	0.769	0.774	0.775	0.775

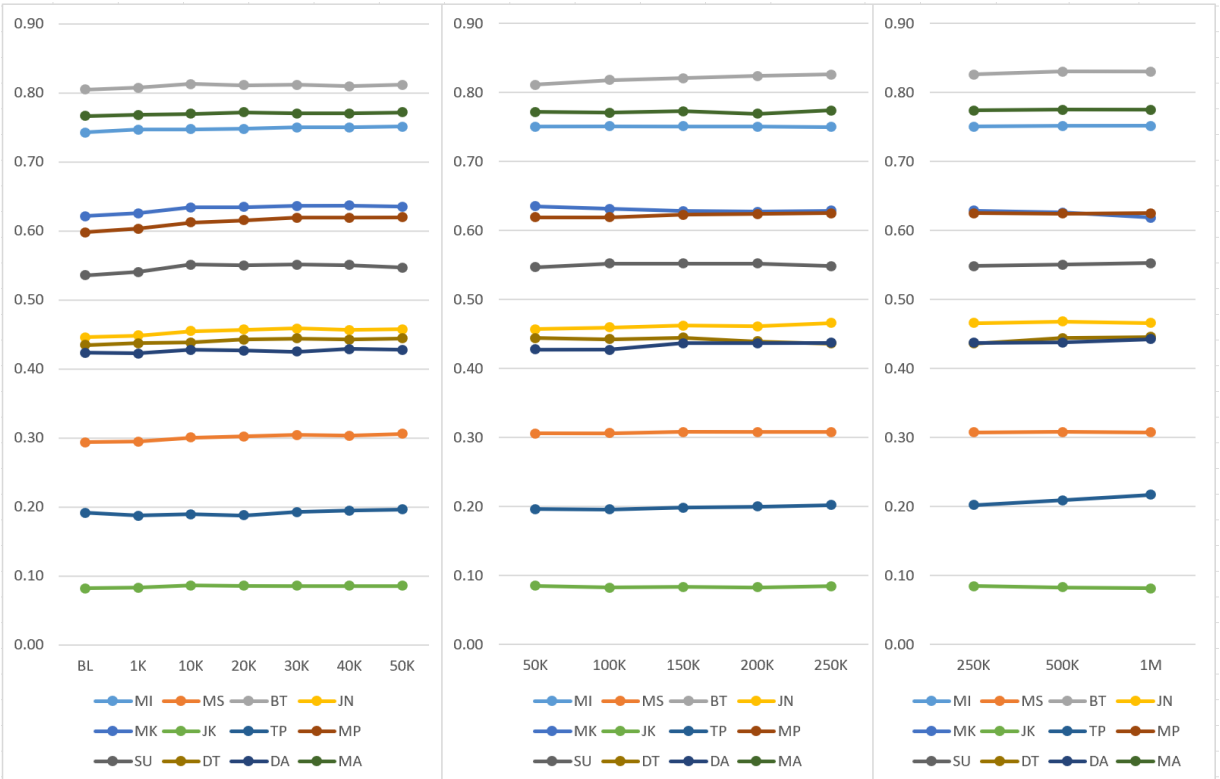

Figure 3. BLEU Points chart with different LM variations.

The results of this study indicate that adding MC to ELRL can increase BLEU points by 0.004 to 0.022 Compared to previous studies in languages with good enough data sources, the increase in BLEU points can reach up to three points. For example, Bojar and Tamchyna [36] describe a series of experiments on English-to-Czech machine translation using a fixed amount of parallel data and varying amounts of monolingual data (ranging from 500,000 to 5 million sentences). The authors evaluated their approach using two separate sets of 1,000 sentences from the CzEng corpus. They found that the gains in the BLEU score become more significant as the size of the monolingual data increases. The most significant improvement in the BLEU score was observed when the monolingual data was largest, resulting in a gain of over three points absolute. The results suggest that incorporating monolingual data can significantly improve machine translation quality, particularly when more data are used.

Fig. 3 displays the dataset's characteristics for each language used in the experiment. By analyzing the pattern of increasing BLEU Points, we can classify them into three different groups. The first group includes languages that consistently improve translation accuracy with increased MC size: BT, MP, and DA. In simpler terms, the machine translation accuracy improves when more data is available in the monolingual corpus for the first group of languages. As for the second group of languages, there is a point at which additional monolingual corpora no longer significantly impact translation accuracy. This group includes languages like MI, MK, JK, SU, DT, and MA. For these languages, the increase in MC up to 50K improves the accuracy of their translation, but adding more MC up to 1M does not significantly improve the accuracy further. The third group consists of languages that show improvement in translation accuracy up to 250K MC but remain relatively constant when additional MCs are added up to 1M. This group includes languages such as MS, JN, and MP. For these languages, the improvement in translation accuracy occurs up to a certain point, but additional MCs do not seem to have a significant impact.

V. Conclusion

In this experiment, we aimed to determine how much of an impact the monolingual corpus's size has on the SMT accuracy in ELRL. The results showed that increasing the quantity of MC, which is used to train the language model on the target side, can improve the accuracy of the machine translation system. However, we found that this improvement is not always consistent, and sometimes adding more MC can decrease accuracy. For instance, we observed a score reduction for some languages when we increased the MC by 200K from 50K to 250K. Therefore, to optimize the accuracy of the SMT system, we need to conduct multiple experiments with different sizes of MCs. By doing so, we can determine the ideal MC quantity required to achieve the best translation results for each language pair.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

H.S. prepared experimental strategies, built models, decoders, and wrote papers; S.C. prepared, cleaned, and analyzed data; A.B.P. evaluated and assisted with writing; all authors have approved the final version.

References

[1] P. Koehn and R. Knowles, "Six challenges for neural machine translation," arXiv preprint, arXiv:1706.03872, Jun. 2017.

[2] R. Östling and J. Tiedemann, "Neural machine translation for low-resource languages," arXiv:1708.05729 [cs.CL], 2017.

[3] M. Dowling, T. Lynn, A. Poncelas, and A. Way, "SMT versus NMT: Preliminary comparisons for Irish," in *Proc. AMTA 2018 Workshop: LoResMT*, 2018.

[4] C. M. Veliz, O. Clercq, and V. Hoste, "Is neural always better? SMT versus NMT for Dutch text normalization," *Expert Syst Appl*, vol. 170, 114500, 2021.

[5] M. Ridwan, "National and official language: The long journey of Indonesian language," *Budapest International Research and Critics Institute-Journal (BIRCI-Journal)*, vol. 1, no. 2, pp. 2615–1715, 2018.

[6] A. C. Cohn and M. Ravindranath, "Local languages in Indonesia: Language maintenance or language shift?" *Masyarakat Linguistik Indonesia*, pp. 131–148, 2014.

[7] D. M. Eberhard, F. S. Gary, and D. F. Charles, *Ethnologue: Languages of the World*, 24th ed. Dallas, Texas: SIL International, 2021.

[8] A. Subiyanto, "Revisiting full reduplication in Indonesian, Javanese, and Sundanese verbs: A distributed reduplication approach," *Culturalistics: Journal of Cultural, Literary, and Linguistic Studies*, vol. 2, no. 2, 2018.

[9] P. Lohar, S. Madden, E. O'Connor, M. Popovic, and T. Habruseva, "Building machine translation system for software product descriptions using domain-specific sub-corpus extraction," in *Proc. the 15th Biennial Conference of the Association for Machine Translation in the Americas*, Sep. 2022, vol. 1, pp. 1–13.

[10] J. Gori, O. Rioul, and Y. Guiard, "Speed-accuracy tradeoff: A Formal Information-Theoretic Transmission Scheme (FITTS)," in *Proc. ACM Transactions on Computer-Human Interaction*, Sep. 2018, vol. 25, no. 5.

[11] T. Mieno, G. Neubig, S. Sakti, T. Toda, and S. Nakamura, "Speed or accuracy? A study in evaluation of simultaneous speech translation," in *Proc. Sixteenth Annual Conference of the International Speech Communication Association*, Sep. 2015, pp. 2267–2271.

[12] D. Torregrosa *et al.*, "Leveraging rule-based machine translation knowledge for under-resourced neural machine translation models," in *Proc. Machine Translation Summit XVII: Translator, Project and User Tracks*, 2019.

[13] P. Koehn, A. Birch, C. Callison-Burch, M. F. Fondazione, and B. Kessler, "Moses: Open source toolkit for statistical machine translation. Marian view project multilingual neural machine translation view project," in *Proc. the 45th Annual Meeting of the Association for Computational Linguistics*, 2007.

[14] Y.-L. Yeong, T.-P. Tan, K. H. Gan, and S. K. Mohammad, "Hybrid machine translation with multi-source encoder-decoder long short-term memory in English-Malay translation," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 8, no. 4-2, pp. 1446–1452, 2018.

[15] A. Karakanta, J. Dehdari, and J. van Genabith, "Neural machine translation for low-resource languages without parallel corpus," *Machine Translation*, vol. 32, no. 1–2, pp. 167–189, Jun. 2018.

[16] R. Sennrich and B. Haddow, "Linguistic input features improve neural machine translation," in *Proc. the First Conference on Machine Translation*, Aug. 2016, vol. 1, pp. 83–91.

[17] M. Johnson *et al.*, "Google's multilingual neural machine translation system: Enabling zero-shot translation," *Trans Assoc Comput. Linguist.*, vol. 5, pp. 339–351, 2017.

[18] T.-V. Ngo, P.-T. Nguyen, T.-L. Ha, K.-Q. Dinh, and L.-M. Nguyen, "Improving multilingual neural machine translation for low-resource languages: French, English-Vietnamese," in *Proc. the 3rd Workshop on Technologies for MT of Low Resource Languages*, Dec. 2020, pp. 55–61.
[19] G. Lample, M. Ott, A. Conneau, L. Denoyer, and M. Ranzato, "Phrase-based & neural unsupervised machine translation," in *Proc. the 2018 Conference on Empirical Methods in Natural Language Processing*, Oct. 2018, pp. 5039–5049.
[20] M. Tars, A. Tättar, and M. Fishel, "Extremely low-resource machine translation for closely related languages," arXiv preprint, arXiv:2105.13065, 2021.
[21] A. L. Tonja, O. Kolesnikova, A. Gelbukh, and G. Sidorov, "Low-resource neural machine translation improvement using source-side monolingual data," *Applied Sciences*, vol. 13, no. 2, 2023.
[22] H. Sujaini, "Improving the role of language model in statistical machine translation (Indonesian-Javanese)," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 2, 2020.
[23] M. Artetxe, G. Labaka, and E. Agirre, "Unsupervised statistical machine translation," in *Proc. the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.
[24] Z. Abidin, Permata, I. Ahmad, and Rusliyawati, "Effect of mono corpus quantity on statistical machine translation Indonesian-Lampung dialect of NYO," *Journal of Physics: Conference Series*, vol. 1751, no. 1, Jan. 2021,
[25] A. E. P. Lesatari, A. Ardiyanti, A. Ardiyanti, I. Asror, and I. Asror, "Phrase based statistical machine translation Javanese-Indonesian," *Jurnal Media Informatika Budidarma*, vol. 5, no. 2, p. 378, 2021.
[26] Q. A. Agigi and A. A. Suryani, "Statistical machine translation Muna to Indonesia language," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 8, no. 4, pp. 2173–2186, 2021
[27] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow, "A survey on recent approaches for natural language processing in low-resource scenarios," in *Proc. the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 2545–2568.
[28] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Commun*, vol. 56, no. 1, pp. 85–100, 2014.
[29] A. Ghafoor *et al.*, "The impact of translating resource-rich datasets to low-resource languages through multilingual text processing," *IEEE Access*, vol. 9, pp. 124478–124490, 2021.
[30] J. Gu, H. Hassan, J. Devlin, V. O. Li, and G. Research, "Universal neural machine translation for extremely low resource languages," in *Proc. NAACL-HLT*, 2018, pp. 344–354.
[31] F. Rahutomo, A. A. Septarina, M. Sarosa, A. Setiawan, and M. M. Huda, "A review on Indonesian machine translation," *Journal of Physics: Conference Series*, 77040, 2019.
[32] S. Cahyawijaya *et al.*, "NusaCrowd: Open source initiative for Indonesian NLP resources," arXiv preprint, arXiv:2212.09648, 2022.
[33] A. A. Suryani, D. H. Widyantoro, A. Purwarianti, and Y. Sudaryat, "Sundanese-Indonesian parallel corpus," *Telkom University Dataverse*, 2022. https://doi.org/10.34820/FK2/HDYWXW
[34] F. Koto and I. Koto, "Towards computational linguistics in Minangkabau language: Studies on sentiment analysis and machine translation," in *Proc. the 34th Pacific Asia Conference on Language, Information and Computation*, Oct. 2020, pp. 138–148.
[35] D. Goldhahn, T. Eckart, and U. Quasthoff, "Building Large monolingual dictionaries at the Leipzig corpus collection: From 100 to 200 languages electronic publishing and multimedia view project frequency dictionaries based on the Leipzig corpus collection," in *Proc. LREC'12*, 2012, pp. 759–765.
[36] O. Bojar and A. Tamchyna, "Improving translation model by monolingual data," in *Proc. the Sixth Workshop on Statistical Machine Translation*, Jul. 2011, pp. 330–336.