Effect of mono corpus quantity on statistical machine translation Indonesian – Lampung dialect of nyo

Z Abidin¹, Permata², I Ahmad³ and Rusliyawati⁴

¹,²,³,⁴ Department of Information System, Faculty of Engineering and Computer Science, University of Teknokrat Indonesia, Jl. ZA Pagar Alam no 9-10, Bandar Lampung, Indonesia
² Department of Informatics, Faculty of Engineering and Computer Science, University of Teknokrat Indonesia, Jl. ZA Pagar Alam no 9-10, Bandar Lampung, Indonesia

email: zabin¹, permata², imamahmad³, rusliyawati⁴@teknokrat.ac.id

Abstract. Lampung Province is located on the island of Sumatera. For the immigrants in Lampung, they have difficulty in communicating with the indigenous people of Lampung. As an alternative, both immigrants and the indigenous people of Lampung speak Indonesian. This research aims to build a language model from Indonesian language and a translation model from the Lampung language dialect of nyo, both models will be combined in a Moses decoder. This research focuses on observing the effect of adding mono corpus to the experimental statistical machine translation of Indonesian - Lampung dialect of nyo. This research uses 3000 pair parallel corpus in Indonesia language and Lampung language dialect of nyo as source language and uses 3000 mono corpus sentences in Lampung language dialect of nyo as target language. The results showed that the accuracy value in bilingual evaluation under-study score when using 1000 sentences, 2000 sentences, 3000 sentences mono corpus show the accuracy value of the bilingual evaluation under-study, respectively, namely 40.97 %, 41.80 % and 45.26 %.

Keyword: mono corpus, language model, translation model, statistical machine translation, bilingual evaluation under-study.

1. Introduction
Lampung Province is a province located at the entrance gate to the island of Sumatera. Lampung Province has rich of culture, one of which is the Lampung language and Lampung script. In general, in Lampung province there are two main dialects, namely the dialect of api and the dialect of nyo. The Lampung provincial government has great concern for the Lampung language. The provincial government continues to make various efforts to preserve and maintain the Lampung language. The Government of Lampung through Governor Regulation number 39 of 2014 concerning Lampung Language and Literacy Subjects stipulates that the Lampung language is a mandatory local content at the primary to senior secondary education unit levels and is supported by the availability of textbooks ranging from elementary, junior high and high school, along with the Lampung language dictionary. The Lampung language, both the dialect of api and the dialect of nyo, is used by the people of
Lampung to communicate daily both in the family environment and at traditional events. The Lampung language belongs to the Austronesian class in the Polynesian Malay language family.

The two main dialects are dialect of api and dialect of nyo which refers to the word "what" [1]. Several research related to the Lampung language have been conducted by researchers. Research on the translation of the Lampung language in dialect of api uses the Neural Machine Translation (NMT) method without the Attention mechanism [2] and the Neural Machine Translation (NMT) method with the Attention mechanism [3]. Research on Statistical Machine Translation (SMT) in Lampung dialect of api sentences [4]. While the Lampung language research from the aspect of speech research was first carried out [5]. Most of the SMT research was carried out at the University of Tanjungpura, including SMT from Bugis Wajo to Indonesian [6] and SMT from Indonesian to Sundanese [7]. Experiment of Indonesian-Japanese SMT [8], Development of Indonesian-Japanese SMT with Lemma Translation and Additional Post Process [9] and English-Bengla SMT [10].

Based on observations of various published manuscripts related to SMT Indonesian to Lampung language dialect of nyo, research on SMT Indonesian into Lampung language dialect of nyo has never been carried out. The research of SMT Indonesian into Lampung dialect of nyo has the potential to be the initial foundation for machine translation for Indonesian to Lampung dialect of nyo. This research aims to build SMT model for the Indonesian - Lampung dialect of Nyo and observe the effect of adding mono corpus on translation accuracy with bilingual evaluation understudy (BLEU) score.

2. Statistical Machine Translation
The main components of SMT are the language model, the translation model and the decoder. The main raw material for SMT research is sentence pairs between the source language and the destination language. In this research, Indonesian as the source language and Lampung language dialect of nyo as the target language.

Source sentences are sentences in certain languages that will be used as input material in the SMT system. In this research, Indonesian language plays a role as the source sentence. Target sentences are sentences in a certain language that will be used as output in the SMT system. In this research, the Lampung language dialect of nyo acts as target sentence. The pair between the source sentence and the target sentence makes corpus parallel. SMT or statistical translation machines in general have an architecture with several components in it, namely the language model, the translation model and the decoder. The language model is used to find the fluency of translation [11]. Translation model serves to find the accuracy (faithfulness) of translation [11]. While the decoder functions to search for text in
the target language that has the highest probability value by considering the language model and the translation model [11].

2.1. Corpus Parallel and Mono Corpus
Lampung language is a language category that is difficult to be used as research material because of the absence of corpus. Efforts made by researchers to be able to meet the need for the existence of parallel corpus and mono corpus is by typing manually. The corpus material is obtained through the Lampung language textbook. The parallel corpus used in this research is 3000 sentences in the Indonesian - Lampung dialect of nyo and the mono corpus used are 3000 sentences in the Lampung language dialect of nyo.

2.2. Component of SMT
Statistical machine translation is a machine that uses statistical approach in processing the translation of a source language into the target language. The statistical approach used is the concept of probability. The concept of probability in a statistical translation machine assumes that each sentence T is a possible translation of the S sentence in the source language [12]. Through the approach that the translated text is based on the P (T | S) probability distribution, it can be done with the Bayes theorem namely:

$$P(T|S) = \frac{P(S|T) P(T)}{P(S)}$$ (1)

The parts that are the key elements in the language model are the probability of a series of words written as $P(w_1, w_2, ..., w_n)$ or $P(w_{1:n})$. The language model assigns probability $P(w_{1:n})$ to a series of n words by averaging a probability distribution. The sequences can be phrases or sentences and the probability can be estimated from a large corpus of documents. One example of a language model approach is the n-gram model. The n-gram language model is a type of probabilistic language model for predicting the next item in the sequence in (n-1) form. The conditional probability can be calculated from the sum of the n-gram frequencies:

$$P(w_i|w_{i-(n-1)}, ..., w_{i-1}) = \frac{\text{count}(w_{i-(n-1)}, w_{i-1}, ..., w_i)}{\text{count}(w_{i-(n-1)}, ..., w_{i-1})}$$ (2)

The following are examples of the n-gram model, namely:
1. Unigram (1-gram) : $P(w_1), P(w_2), ..., P(w_n)$.
2. Bigram (2-gram) : $P(w_1), P(w_2|w_1), ..., P(w_n|w_{n-1})$.
3. Trigram (3-gram) : $P(w_{1:n}) = P(w_1), P(w_2|w_1), P(w_3|w_{1:2}) ... P(w_{n-2}|w_{n-1})$

The translation model is used to match the input text in the source language with the output text in the target language. In statistical translation machine there are two translation models, namely the word-based translation model (word-based translation model) and phrase-based translation model (phrase-based translation model) [12]. The decoder is in charge of finding text in the target language that has the greatest probability by considering the translation model and language model factors [12]. The calculation for $\hat{T}$ (translation result) can be written as follows:

$$\hat{T} = \arg_T \max P(T|S) = \arg_T \max \frac{P(S|T)}{P(S)} P(T) = \arg_T \max P(S|T).P(T)$$ (3)

2.3. Automatic Evaluation
The evaluation of the translation results is done by comparing the translated sentences with the reference sentences using the Bilingual Evaluation Understudy (BLEU). BLEU is an algorithm that functions to evaluate the quality of the translation results that have been translated by machines from a source language to a destination language. BLEU measures a statistically-based precision score that
has been modified between the translation results, automatically, with a reference translation using a constant called a brevity penalty (BP) [13].

\[
BP_{\text{BLEU}} = \begin{cases} 
1 & \text{jika } c > r \\
\exp(1 - \frac{c}{r}) & \text{jika } c \leq r
\end{cases}
\]

\[p_n = \frac{\sum_{C \in \text{Candidates}} \sum_{n \text{ gram} \in C} \text{count}_{\text{clip}}(n \text{ gram})}{\sum_{C \in \text{Candidates}} \sum_{n \text{ gram} \in C} \text{count}_{\text{clip}}(n \text{ gram})} \]

\[
\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \cdot \log p_n\right)
\]

BP is a symbol of brevity penalty, \(c\) is the number of words from the machine translation result, \(r\) is the length of the effective reference translation. \(P_n\) is the geometric mean of the modified \(n\)-gram precision. N values used are \(N = 4\) and \(w_n = \frac{1}{N}\) [14].

3. Research Methodology
The research methodology carried out is depicted in Figure 2 below:

![Figure 2. Research Flow Diagram](image)

3.1. Data Collection
The data used in this research are material documents in the Lampung language dialect of nyo which is translated into Indonesian. The documents are taken from the Lampung language book for elementary and junior high school levels in the province of Lampung. Document data is processed into text data in * .lg and * .id formats. The document data then will be processed into a parallel corpus in the Indonesian language - Lampung in the dialect of nyo and the mono corpus for the Lampung language in the dialect of nyo.
3.2. **Parallel and Mono Corpus Creation**

The document data that has been collected will then be manually typed to be made into a parallel corpus in Indonesian – Lampung dialect of nyo and mono corpus in Lampung dialect of nyo. There were 3000 parallel corpus sentences collected in Indonesian - Lampung dialect of nyo and 3000 mono corpus sentences in Indonesian - Lampung dialect of nyo. The corpus is saved in *.lg and *.id formats.

3.3. **Implementation of SMT**

The architecture of SMT Indonesian - Lampung dialect of nyo built in this research consists of (1) parallel corpus of 3000 pairs of Indonesian - Lampung dialect of nyo sentences, (2) 3000 sentences of mono corpus of Lampung in dialect of nyo (3) generating a language model with GIZA ++, GIZA ++ for word-aligning our parallel copus [15], (4) generating a translation model with KenLM, KenLM is included in Moses and the default in the Moses tool-chain [15]. The language model (LM) is used to ensure fluent output, so it is built with the target language. The KenLM documentation gives a full explanation of the commandline options, but the following will build an appropriate 3-gram language model [15]. (5) combining language models and translation models with Moses Decoder. The job of the Moses decoder is to find the highest scoring sentence in the target language (according to the translation model) corresponding to a given source sentence [15]. (6) conducting trials with Moses, (7) Assessing the translation results through the BLEU score.

![Figure 3. Architecture of SMT Indonesian-Lampung dialect of nyo](image)

Experiments in the research of SMT in Indonesian - Lampung dialect of nyo were carried out three times with the following conditions:

1. The first experiment was carried out with the conditions of 3000 pairs of sentences in Indonesian - Lampung dialect of nyo and 1000 mono corpus sentences in Lampung dialect of nyo.
2. The second experiment was carried out with the conditions of 3000 pairs of sentences in Indonesian - Lampung dialect of nyo and 2000 mono corpus sentences in Lampung dialect of nyo.
3. The third experiment was carried out with the conditions of 3000 pairs of sentences in Indonesian-Lampung dialect of nyo and 3000 mono corpus sentences in Lampung dialect of nyo.

Tests were performed on each experiment using 100 Indonesian new sentences that are not available in 3000 Indonesian languages-Lampung dialect of nyo and 3000 mono corpus sentences Lampung dialect of nyo. Evaluation of the results of each experiment is seen through the BLEU score. The research material in the form of 3000 sentence pairs in Indonesian-Lampung dialect of nyo and 3000 sentences of mono corpus in Lampung dialect of nyo and 100 test sentences in Indonesian, the syntax used for SMT research is also provided at the link or can be accessed at the following link: [https://drive.google.com/drive/folders/1oNpybrq5OJ4Ne0HS5w9eHqnZIZASpmC?usp=sharing](https://drive.google.com/drive/folders/1oNpybrq5OJ4Ne0HS5w9eHqnZIZASpmC?usp=sharing).

4. Experimental Results

The SMT experiment was carried out three times to observe the effect of adding the mono corpus quantity on the translation of Indonesian into the Lampung language dialect of nyo. At the end of each experiment the BLEU value and the translation results will be observed.

4.1. Implementation of SMT Indonesian – Lampung dialect of nyo

The pre-processing phase of SMT of the Moses Decoder consists of several parts, namely sentence alignment, tokenization, cleaning, lowercase filtering, and truecase. Sentence alignment is the process of aligning the parallel corpus of Indonesian-Lampung dialect of nyo. Tokenization is needed to provide spacing between words, including spacing between words and existing punctuation marks, while lowercase is a process to uniform the capitalization of letters. The commands used in tokenization at Moses are shown in figure 4 below.

**Figure 4. Syntax sample for tokenization and lower case filtering [15]**

In this truecasing process, each beginning of each sentence is converted to the most likely place. The commands used in truecasing at Moses are given in figure 5 below.

**Figure 5. Syntax sample for truecasing and lower case filtering**
Cleaning is the process of giving a sentence length limit. Cleaning also works to remove inconsistent sentences. The command used in the cleaning in Moses is shown in figure 6 below.

Figure 5. Syntax sample for truecasing [15]

The next phase is the training phase. It is in this phase that the language model and model translation are carried out. Language model using software, in this study used KenLM, which has been integrated into Moses. The commands used in the language model in Moses are given in Figure 7 below.

Figure 6. Syntax sample for cleaning [15]

Figure 7. Syntax sample for training phase [15]

Giza ++ software is used for translation model which implements the Expectation Maximization algorithm or the IBM model algorithm. The commands used in training at Moses are given in Figure 8 below.

Figure 8. Syntax sample for translation model [15]

Especially for the translation model, the approach used is phrase-based, in which there are several stages. This stage starts from the alignment of words to the creation of a configuration file that will be used in the decoding process. The testing phase is the phase in translating the source language into the target language. Input in the form of a sentence in the source language will be translated into the target language. This phase is also the decoding phase of the results of the training phase. The commands used in tuning in Moses are given in figure 9 below.
The last phase is the automatic evaluation phase of the translation results obtained from the testing phase. The instructions used in testing in Moses are given in figure 10 below.

Figures 4 to 10 above show the syntax for the first experiment. The second and third experiments used the same method except for a different number of parallel corpuses.
4.2. Testing of SMT Indonesian – Lampung dialect of nyo

Testing of automatic translation results is carried out using 100 sentences that are not in the parallel corpus based on the BLEU value. The model that has been built uses the Moses Decoder from 3000 sentences of parallel corpus pairs. The results of the Indonesian SMT Experiment - Lampung Dialect of Nyo can be seen in table 1 and figure 11 below.

| Experiment                              | BLEU (%) |
|-----------------------------------------|----------|
| 1000 sentences mono corpus              | 40.97    |
| 2000 sentences mono corpus              | 41.8     |
| 3000 sentences mono corpus              | 45.26    |

Figure 11. Results of SMT Indonesian-Lampung Dialect of nyo

4.3. Analysis of test results

Analysis of the test results was carried out to determine the characteristics of the machine translator and identify whether the results were in accordance with the needs based on the results of the accuracy test of the SMT Indonesian-Lampung dialect of nyo.

| Test Sentences in Indonesian | Reference Sentences in Lampung Dialect of nyo | The results of first experiment | The results of second experiment | The results of third experiment |
|------------------------------|-----------------------------------------------|--------------------------------|----------------------------------|--------------------------------|
|                              |                                               |                                |                                  |                                |
Information from table 2 shows five samples of the test sentences used in this research. The first column in table 2 is the test sentence in Indonesian language, the second column is the reference sentence for the translation of the test sentence in Lampung dialect of nyo, the third column is the translation result of the test sentence in the first experiment, the fourth column is the translation result of the test sentence in the second experiment and the fifth column is the translation result of the test sentence in the third experiment.

In the first sample sentence, all experiments show the same translation results. In the second sample sentence, the word ‘bertengkar’ was not translated successfully in all experiments. However, in the first experiment the SMT system gave the result that the translation was not as expected and word ‘bertengkar’ was considered as out-of-vocabulary. In the second and third experiments in the second sample sentence, the translation results are almost the same as in the first experiment except for the position of the word 'laho'. In the third sample sentence, all experiments show the same translation results. The word ‘bertemu’ in Indonesian is translated as the word ‘tumbuk’, whereas it should have been ‘tunggo’. In the fourth sample sentence, the results of the third experiment show better translation results and the first experiment and the second. In the fifth sample sentence, all experiments show the same translation results.

5. Conclusion and future work
In this research, the translation of Indonesian into Lampung in the nyo dialect can be done using the Moses Decoder. This research focuses on observing the effect of adding mono corpus. From the experimental results, it can be seen that the addition of a mono corpus contributes to an increase in the accuracy of the translation results. When using 1000 mono corpus, the BLEU value was 40.97%, 2000 mono corpus obtained a BLEU value of 41.80%, while 3000 mono corpus obtained a BLEU value of 45.26%. This research still has the potential to be further developed in the aspect of investigating the quality of the corpus used, adding POS tags to the target language.

Acknowledgments
This publication is the output of a beginner lecturer research grant from the Ministry of Research and Technology / National Agency for Research and Innovation (Kemenristek / BRIN) with contract number 078/SP2H/LT/DRPM/2020, 839/SP2H/LT/MONO/LL2/2020, 046/UTI/LPPM/E.1.3/VII/2020..
References

[1] F. Ariyani, “Distribusi Verba Berrefiks (N-) Pada Bahasa Lampung dalam Kitab Kuntara Raja Niti dan Buku Ajar. Ranah: Jurnal Kajian Bahasa 3,” Ranah J. Kaji. Bhxs., vol. 3, no. 2, pp. 124–134, 2014, doi: https://doi.org/10.26499/rmh.v3i2.43.

[2] Z. Abidin, “Penerapan Neural Machine Translation untuk Eksperimen Penerjemahan secara Otomatis pada Bahasa Lampung – Indonesia,” Pros. Semin. Nas. Metod. Kuantitatif 2017, no. 978, pp. 53–68, 2017.

[3] Z. Abidin, A. Sucipto, and A. Budiman, “Penerjemahan Kalimat Bahasa Lampung-Indonesia Dengan Pendekatan Neural Machine Translation Berbasis Attention Translation of Sentence Lampung-Indonesian Languages With Neural Machine Translation Attention Based,” J. Kelitbangan, vol. 06, no. 02, pp. 191–206, 2018.

[4] P. Permata and Z. Abidin, “Statistical Machine Translation Pada Bahasa Lampung Dialek Api Ke Bahasa Indonesia,” Media Inform. Budidarma, vol. 4, no. 3, pp. 519–528, 2020, doi: 10.30865/mib.v4i3.2116.

[5] S. Ningsih and S. Saniati, “Eksperimen Pengenalan Ucapan Aksara Lampung Dengan CMU Sphinx 4,” J. Teknoinfo, vol. 12, no. 1, p. 33, 2018, doi: 10.33365/jti.v12i1.40.

[6] T. Apriani, H. Sujaiani, and N. Safriadi, “Pengaruh Kuantitas Korpus Terhadap Akurasi Mesin Penerjemah Statistik Bahasa Bugis Wajo Ke Bahasa Indonesia,” J. Sist. dan Teknol. Inf., vol. 1, no. 1, pp. 1–6, 2016.

[7] R. Darwis, H. Sujaiani, and R. D. Nyoto, “Peningkatan Mesin Penerjemah Statistik dengan Menambah Kuantitas Korpus Monolingual (Studi Kasus : Bahasa Indonesia - Sunda),” J. Sist. dan Teknol. Inf., vol. 7, no. 1, p. 27, 2019, doi: 10.26418/justin.v7i1.27254.

[8] H. S. Simon and A. Purwarianti, “Experiments on Indonesian-Japanese statistical machine translation,” Proceeding - IEEE Cybern. 2013 IEEE Int. Conf. Comput. Intell. Cybern., pp. 80–84, 2013, doi: 10.1109/CyberneticsCom.2013.6865786.

[9] M. A. Sulaeman and A. Purwarianti, “Development of Indonesian-Japanese statistical machine translation using lemma translation and additional post-process,” Proc. - 5th Int. Conf. Electr. Eng. Informatics Bridg. Knowl. between Acad. Ind. Community, ICEEI 2015, no. i, pp. 54–58, 2015, doi: 10.1109/ICEEI.2015.7352469.

[10] A. H. Imam, M. R. Mahmud Arman, S. H. Chowdhury, and K. Mahmood, “Impact of corpus size and quality on English-Bangla statistical machine translation system,” 14th Int. Conf. Comput. Inf. Technol. ICCT 2011, no. Iccit, pp. 566–571, 2011, doi: 10.1109/ICCITechn.2011.6164853.

[11] R. Nugroho Aditya, T. Adj Valiant, and B. Hantono S, “Penerjemah Bahasa Indonesia dan Bahasa Jawa Menggunakan Metode Statistik Benzah Soner,” Semin. Nas. Teknol. Inf. dan Komun., vol. 2015, no. Sentika, 2015.

[12] H. Tanuwijaya and H. Manurung, Maruli, “Penerjemah Dokumen Inggris-Indonesia Menggunakan Mesin Penerjemah Statistik Dengan Word Reordering dan Phrase Reordering,” J. Ilmu Komput. dan Inf., vol. 2, no. 1, pp. 17–24, 2009.

[13] A. Hermanto, T. Adj, and N. A. Setiawan, “Recurrent neural network language model for English-Indonesian Machine Translation: Experimental study,,” in ICSITech, 2015, pp. 132–136.

[14] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a Method for Automatic Evaluation of Machine Translation,” Proc. 40th Annu. Meet. Assoc. Comput. Linguist., pp. 311–318, 2002, doi: 10.1002/andp.19923712302.

[15] P. Koehn, Statistical Machine Translation SystemUser Manual and Code Guide. 2019.