HASIL CEK_Dewi Soyusiawaty, Bella Okta Sari Miranda

by Dewi Soyusiawaty, Bella Okta Sari Miranda Statistical Machine Translation From Indonesian To

Submission date: 30-Jan-2023 11:26AM (UTC+0700)
Submission ID: 2002168202
File name: SMT-IJCA_R1.doc (326.5K)
Word count: 3973
Character count: 21558

Statistical Machine Translation From Indonesian to Regional Languages in Indonesia

Dewi Soyusiawaty
Informatics, Universitas Ahmad Dahlan
Kampus 4, Ringroad selatan, Yogyakarta
dewi.soyusiawaty@tif.uad.ac.id

Bella Okta Sari Miranda
Informatics, Universitas Ahmad Dahlan
Kampus 4, Ringroad selatan, Yogyakarta
mirandabella1110@gmail.com

ABSTRACT
The current condition in Indonesia has 617 regional languages. There are 15 regional languages that are declared extinct and 139 others are in endangered status. Utilization of computer-based tools can be used as an effort to preserve regional languages digitally according to current technological developments, including by building digital dictionaries and translation machines. The digital dictionary has the ability to translate regional languages into Indonesian with the approach used is translating word for word, although it is not effective when done manually. An alternative solution is to create a machine translation application. Machine translation can be dictionary-based or language-parallel corpus data-based. Statistical Machine Translation (SMT) is a machine translation approach with translation results generated on the basis of a statistical model whose parameters are taken from the results of a parallel corpus analysis. The quality of the SMT translation results is influenced by several factors. The most fundamental factor is the number of parallel corpus available and the quality of the corpus used as the basis for building translation models and language models. This study aims to determine the role of parallel corpus in improving SMT accuracy, especially in regional languages in Indonesia. The research data used is parallel corpus text of 3000 pairs of sentences. Based on the results of the research that has been done, it is found that the optimization of parallel corpus can increase the value of translation accuracy. Better translation accuracy can be achieved with optimized parallel corpus. Besides that, testing with single sentences will provide higher accuracy than using compound sentences. Testing of 3000 random parallel corpus parallels can increase accuracy by 11.4%, higher than testing with 3000 random parallel corpus.

General Terms
Natural Language Processing

Keywords
Regional languages, Bengkulu Malay, BLEU, Parallel Corpus, Statistical Machine Translation

1. INTRODUCTION
The Indonesian nation has a diverse culture, including regional languages in it. To understand and explore the culture of a nation, the most effective way is to first learn and know the language used by that nation. The diversity of languages in Indonesia is actually capable of enriching Indonesian culture, but in reality the existence of regional languages is starting to be threatened. Current conditions in Indonesia, out of 617 regional languages, 15 regional languages have been identified as extinct and 139 others are in endangered status. Efforts to maintain and preserve regional languages must not only be carried out by the central government, but also require the active role of local governments, communities, and educational institutions or educational institutions.[1][2]

There are many ways that can be done as an effort to preserve the language culture of a region carried out by the government, for example by regional television programs broadcasting news broadcasts in regional languages, playing programs showing stories for children and adults in regional languages. Utilization of computer-based tools can be used as an effort to preserve regional languages digitally according to current technological developments, including by building digital dictionaries and translation machines. The digital dictionary has the ability to translate regional languages into Indonesian with the approach used is translating word for word, although it is not effective when done manually. An alternative solution is to create a machine translation application. Machine translation can be dictionary-based or language-parallel corpus data-based. For machine translation, one of the machine translation tools currently available is Google translator. Currently on Google's translation engine there are only two local languages available, namely Javanese and Sundanese.[3]

There are many similarities between regional languages and Indonesian. For example, in Bengkulu Malay, there is a change in basic word patterns such as "saya" becomes "saye" (the ending -a becomes -e), "bayar" becomes "baya" (the ending -ar becomes -a), "kecil" becomes "kecik " (the suffix -il becomes -ik), "balik" becomes "balek" (the suffix -ik becomes -ek), "bangkit" becomes "bangket" (the suffix -it becomes -et), "buih" becomes "bueh" (the ending –ih becomes –eh) and so on. Likewise with other regional languages, namely Acehnese, Riau Malay, Minang languages, etc. The sentence order in Indonesian is the same as the word order in the regional target language. Characteristics of local languages like this is a potential for developing Rule Based Machine Translation (RBMT).[4][5]

Rule-Based Machine Translation involves morphological, syntactic, and semantic rules about the source and target language. This system can handle word-order problems. RBMT process the system word by word and can't handle ambiguity and idiomatic expression. Hence the resulting translation often not fluent and can't generate natural translation. Another weakness in RBMT is the large number of rules that must be made.[6]

Statistical Machine Translation (SMT) is a machine translation approach where translation results are generated based on a statistical model whose parameters are taken from the analysis of a parallel corpus, or commonly called a bilingual corpus, from two different languages. SMT can handle morphology because it can separate suffixes that inflected word leading to meaning transfer. In other words, SMT can handle ambiguity. The system records phrase-based translations with their frequency of occurrence on phrase table. Thus, the translation result generates more fluent and natural than RBMT. One weakness of SMT is the challenge of translating material that is not similar to content from the training corpora. It gives poor accuracy of the translation result. So that, to achieved good translation, the corpus should be customized for a specific style. SMT does not work well between languages that have significantly different word orders e.g. Japanese-Indonesian.[3] [7]

Research in the field of SMT in Indonesia, especially for machine translation of Indonesian into regional languages, has begun to be carried out a lot, including research on the accuracy of Indonesian-Javanese translation using phrase-based statistical methods, research on the effect of corpus quantity on the MPS accuracy of Bugis Wajo to Indonesian, research by improving the probability of the Lexical Model to improve the accuracy of Indonesian-Javanese MPS, Marking base words and affixes in the parallel corpus to improve the accuracy of Indonesian–Dayak Taman translations, tuning for quality research to test the accuracy of Indonesian-Dayak Kanayatn MPS, research on the Javanese-Javanese script translation system based on finite state automata and Statistical Machine Translation Dayak Language – Indonesia language.[8][9]

Training the translational model components in SMT requires a large parallel corpus for the parameters to be predicted. Therefore, higher translation accuracy can be achieved when the machine translation system is trained to increase the number of parallel corpus. The output quality of the SMT system is highly dependent on the quality of the sentence pairs. Manual compilation of parallel corpus is too expensive, so most of the available parallel corpus is generated automatically. [9]Automated methods for constructing parallel pairs are largely imprecise and using a low-quality training corpus that has many nonparallel pairs will result in low-quality translations. Errors in the automatically generated corpus may be caused by differences between source and target document content, non-literal translations or sentence alignment errors. It is nearly impossible to eliminate alignment errors manually in a large parallel corpus.[10][11]

2. LITERATURE REVIEW

2.1 Statistical Machine Translation

Statistical machine translation is a type of machine translation that uses a statistical approach. The statistical approach used is the concept of probability. Each pair of statements (S,T) is assigned a P(T|S). This can be interpreted as the probability distribution that the interpreter will produce T in the target language given S in the source language. As shown in Figure 1, the process of translating a sentence from one language to another involves her three components: a language model, a translation model, and a decoder. [12]

a) Language Model (LM)

Language Model functions to counting probability from the sentences that possibly appear. Before counting the sentences probabilities, firstly need to count the word probabilities as word order comes first than sentence order with chain rule or commonly called n-gram model formula.

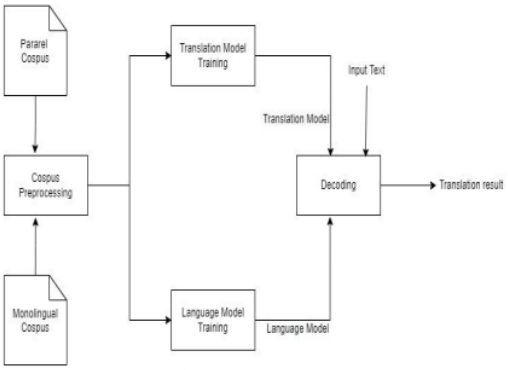

Fig 1 : SMT Components

The N-gram language model consists of three languages: Unigrams, bigrams and trigrams. Unigrams are occurrences of words that are not influenced by other words. A bigram is the occurrence of a word influenced by another word. A trigram is an occurrence of a word influenced by a previous word. Followed by three N-gram language models. [13]

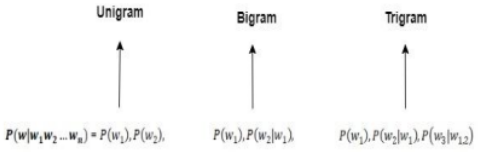

Fig 2 : N-gram language model

b) Translation Model (TM)

A translation model is used to pair the input text in the source language with the output text in the target language. Statistical machine translation has his two translation models: word-based translation model and sentence-based translation model.

c) Parallel Corpus

A collection of texts aligned or harmonized with two or more other languages. Parallel corpora are applied to theoretical problems such as studying translation processes and how ideas are expressed in two (or more) different languages, or comparing properties of original (source) and translated texts. can. Parallel corpora are also useful for other natural language processing applications such as: cross-language information retrieval, word disambiguation, and annotation projection. Building a corpus containing a large number of parallel sets is one of the most time-consuming and critical tasks for high-performance SMT systems. [14]

1.3 Evaluation Method

BLEU (Bilingual Evaluation Understudy) is an algorithm used to evaluate the quality of machine translations from one natural language to another. BLEU measures the modified n-gram accuracy score between the results of automatic

translation and reference translation using a constant called brevity penalty [8]. The BLEU value is obtained by multiplying the shortening penalty by the geometric mean of the corrected precision values. To achieve a high BLEU score, the length of the translation should be close to the length of the reference, and the translation should have the same words and order as the reference [8]. The BLEU formula is as follows:

$$BP = \begin{cases} 1, & c > r \\ e^{\left(1-\frac{r}{c}\right)}, & c \leq r \end{cases}$$

$$\text{BLEU} = \text{BP} \times exp\left(\sum_{n=1}^{N} Wn \log Pn\right)$$

BP = brevity penalty

c = number of words from the automatic translation result

r = number of reference words

Wn = 1/N (standard value of N for BLEU is 4)

Pn = number of n-grams of translation results that are in accordance with the reference divided by the number of n-grams of translation results

3. METHOD

3.1 Corpus Text

The data used in this study are Bengkulu Malay language text documents sourced from the Bengkulu Malay language dictionary. Computationally, this research requires materials in the form of Bengkulu Malay sentences which already have meaning in Indonesian. This language was chosen to represent several regional languages in Indonesia. The general characteristic of some of these regional languages is that they have the same word order between Indonesian sentences and their translation into regional languages, then there is no change in the number of words between compound words in Indonesian and compound words in regional languages. Sentences are typed manually to create a parallel corpus. The document consists of 3000 pairs of Indonesian-Malay Bengkulu sentences. Parallel corpus is stored in .id format for Indonesian language corpus and in .bkl format for Bengkulu Malay language corpus. The following are examples of Indonesian sentences and their translations in Bengkulu Malay.

Table 1. Example of Parallel Corpus Indonesian – Bengkulu Malay

No	Indonesian Sentences	Bengkulu Malay Sentences
1	Siapa yang adzan subuh tadi suaranya sangat bagus	siapo yang bang subuh tadi suaronyo elok nian
2	Bapak Ibu saya telah meninggal semua	bak mak ambo lah meninggal segalo nyo
3	Saya membuat rumah ini dengan cara berangsur-angsur	ambo membuek rumah iko dengan caro berguyur
5	Di rumah ini sampah berserakan dimana-mana	di rumah iko sarok berceceran dimano-mano
6	Kamu jangan hanya bisa mencela orang saja	kau ko jangan pacak mencacek orang ajo

Table 2. Examples of Sentences in Indonesian and Bengkulu Malay

Bahasa Indonesia	Siapa	adzan	subuh	tadi	suaranya	bagus	sekali
Bahasa Melayu Bengkulu	siapo	bang	subuh	tadi	suaronyo	elok	nian

Bahasa Indonesia	Bapak	Ibu	saya	telah	meninggal	semua
Bahasa Melayu Bengkulu	bak	mak	ambo	lah	meninggal	segalonyo

3.2 Design of Statistical Machine Translation

The architecture of the statistical translation machine in the following figure:

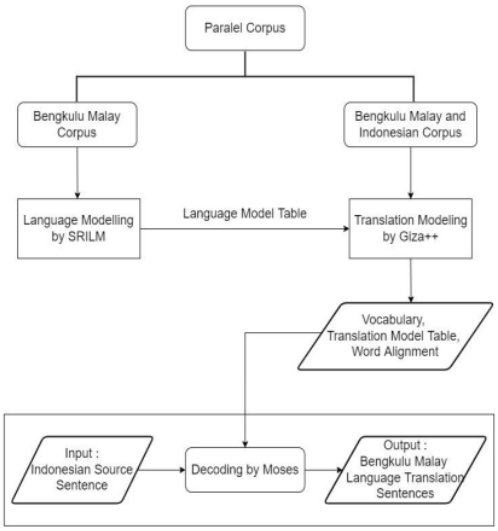

Fig 3 : SMT architecture

Starting with the creation of a parallel corpus followed by the modeling stage in the target language and modeling the translation of the source language into the target language. The language model is used to ensure good output, so that it is built with the target language. Then the stages of the decoding process are combining the language model and the translation model. The decoder finds the sentence with the highest score in the target language (according to the translation model) that matches the given source sentence, then ends with the evaluation stage, namely assessing the translation results through a BLEU score.

3.3 System Testing

The test is carried out by comparing the corpus with the translation data of the source sentences and target sentences whose contents are in accordance with each other (optimized parallel corpus) with the corpus whose contents are random

(random parallel corpus). A random parallel corpus is a corpus where the sentence order between the source sentence and the target sentence is not the same and there are also spelling errors between the source sentence and the target sentence. The test was carried out five times using 3000 random parallel corpus sentences and 3000 optimized parallel corpus sentences. In each test carried out with 600 test sentences.

4. RESULT AND DISCUSSION

a. Cleaning

Cleaning is done by removing empty sentences and removing excessive space characters and removing long sentences because they can cause problems with the training flow, and sentences that are clearly out of tune will be deleted. Cleaning results can be seen in table 3.

Table 3. Cleaning

No	Indonesian	Bengkulu Malay
1	Siapa yang mengatakan saya lelah?	Siapo yang ngecek ambo litak?
2	Bagaimana biar masak enak?	Cakmano mangko masak lemak?
3	Lelah saya ini	Litak ambo ko
4	Warung ibu saya sangat sepi	Lepau mak ambo sepi nian
5	Coba lihat lalat besar itu	Cubo tengok langau besak tu

b. Lowercase

The lowercase stage is to equalize all letters in lowercase without changing the word structure. Lowercase results can be seen in table 4.

Table 4. Lowercase

No	Indonesian	Bengkulu Malay
1	siapa yang mengatakan saya lelah?	siapo yang ngecek ambo litak?
2	bagaimana biar masak enak?	cakmano mangko masak lemak?
3	lelah saya ini	litak ambo ko
4	warung ibu saya sangat sepi	lepau mak ambo sepi nian
5	coba lihat lalat besar itu	cubo tengok langau besak tu

c. Tokenization

Tokenization broadly breaks a set of characters into words. This step will remove punctuation marks, numbers and other characters that are not in the alphabet so that the word will stand alone. The results of tokenization can be seen in table 5.

Table 5. Lowercase

No	Indonesian		Bengkulu Malay	
	Sentences	Tokenizing	Sentences	Tokenizing
1	siapa yang mengatakan saya lelah	siapa	siapo yang ngecek ambo litak	siapo
		yang		yang
		mengatakan		ngecek
		saya		ambo
		lelah		litak
2	bagaimana biar masak enak	bagaimana	cakmano mangko masak lemak	cakmano
		biar		mangko
		masak		masak
		enak		lemak
3	lelah saya	lelah	litak	litak
	ini	saya	ambo ko	ambo
		ini		ko
4	warung ibu saya sangat sepi	warung	lepau mak ambo sepi nian	lepau
		ibu		mak
		saya		ambo
		sangat		sepi
		sepi		nian
5	coba lihat lalat besar itu	coba	cubo tengok langau besak tu	cubo
		lihat		tengok
		lalat		langau
		besar		besak
		itu		tu

d. Training Phase

The next phase is the training phase. In this phase, the language model and translation model are carried out.

Language modeling is carried out to obtain a language model of the target language, namely Bengkulu Malay. The language model is used as a source of knowledge or a source of text-based information with probability values. The language model used in this research is n-gram data and the language model built with SRILM tools produces output in the bkl.lm file format. The language model table produced by SRILM can be seen in Figure 4 below.

```
\data\
ngram 1=2222
ngram 2=10302
ngram 3=1201

\1-grams:
-4.30298   ada       -0.1475204
-2.855822 abang     -1.157007
-3.60401   acara     -0.07446224
-----------------------------------------------------------------------
-------------
\2-grams:
-0.9160199          <s> ambo  -0.1294511
-2.014723 <s> anak  0.09438016
-2.994077 <s> badan 0.104811
-----------------------------------------------------------------------
-------------
\3-grams:
-0.4220276          dekek abang kau
-0.6084259          kek abang ambo
-0.4220276          ketemu abang kau
```

Fig 4 : The Bengkulu Malay-Indonesian language model

The language model generates data ngrams consisting of n gram 1, n gram 2, n gram 3. Unigram (n gram 1) has one token data, bigram (n gram 2) has two token data and trigram (n gram 3) has data three tokens and each data from n gram is included with its probability value.

e. Translation Model

a translation model built from a parallel corpus of Indonesian – Bengkulu Malay using Giza++ . The translation model functions to pair the input text from Indonesian to the output text from Bengkulu Malay. The results of the translation model are in the form of vocabulary corpus document files, word alignment and translation model tables. An example of vocabulary is shown in Figure 5.

```
1     UNK     0
2     ambo    1209
3     kau     780
4     ?       579
5     nyo     406
6     idak    373
7     nian    352
8     ko      332
9     iko     283
10    yang    205
11    apo     168
```

Fig 5 : Example of Bengkulu Malay corpus vocabulary

The numbers 1 to 15 in the corpus vocabulary document are the unique id for each data token, while the numbers to the right of the token indicate the frequency of occurrence of each word. The vocabulary generated by the Indonesian – Bengkulu Malay translation machine consists of 2028 tokens for the Indonesian corpus and 2220 tokens for Bengkulu Malay.

The word alignment document for Indonesian–Bengkulu Malay has 3 lines of sentences. The first line contains the location of the target sentence (1) in the corpus, the length of the source sentence (8), the length of the target sentence (9) and the alignment score. The second line is the source language and the third line is the alignment of the target language sentence to the source language sentence. The word "galo" ({ 5 }) means that the word "galo" in the target language sentence is aligned to the fifth word in the source language sentence, namely "semua". An example of a word alignment document is shown in Figure 6 below.

```
# Sentence pair (1) source length 9 target length 8 alignment
score : 2.40773e-08

bunga dihalaman rumah dibersihkan semua oleh tukang
rumput

NULL ({ }) bungo ({ 1 }) dilaman ({ 2 }) rumah ({ 3 }) di ({ 4 })
babat ({ 6 }) galo ({ 5 }) dekek ({ }) tukang ({ 7 }) rumput ({ 8 })
.....
```

Fig 6 : Alignment document Indonesian - Bengkulu Malay

f. Testing

Automatic testing of machine translation produces output in the form of accuracy values generated using BLEU (Billingual Evaluation Understudy) on the Moses Decoder. Testing is carried out by using a random parallel corpus, or corpus before optimization and using the optimized parallel corpus.

1. Testing with 3000 random parallel corpus

The test was carried out five times. In each test carried out with 600 test sentences. SMT test results can be seen in table 6. The average BLEU score is 84.07%.

Table 6. Accuracy Testing 3000 random parallel corpus

Test group	Corpus (*Fold*)	Test sentence (*fold*)	BLEU(%)
A	2,3,4,5	1	82.28
B	1,3,4,5	2	85.41
C	1,2,4,5	3	86.79
D	1,2,3,5	4	88.83
E	1,2,3,4	5	77.02
		Average	84.07

2. Testing with 3000 optimized parallel corpus

BLEU average for testing of 3000 optimized parallel corpus is 95.47%. Optimization of the parallel corpus increases the accuracy of the Indonesian-Malay Bengkulu translation by 11.4%. The results of the BLEU value can be seen in table 7.

Table 7. Accuracy Testing 3000 random parallel corpus

Test group	Corpus (*Fold*)	Test sentence (*fold*)	BLEU(%)
A	2,3,4,5	1	93.48
B	1,3,4,5	2	95.06
C	1,2,4,5	3	95.56
D	1,2,3,5	4	96.48
E	1,2,3,4	5	96.77
		Average	95.47

A comparison graph of BLEU scores for testing of 3000 random parallel corpus and 3000 optimized parallel corpus is presented in Figure 7.

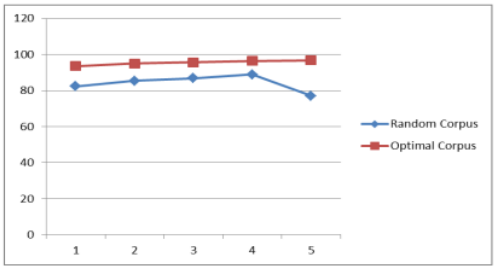

Fig 7 : Comparison of BLEU score on testing 3000 random parallel bodies and 3000 optimized parallel corpus

There are several examples of the results of statistical machine translation in Indonesian-Malay Bengkulu which show that optimized parallel corpus has succeeded in correcting translation errors compared with random parallel corpus. Example," dia memberi makan ikan di kolam paman" is translated by using random parallel corpus (RPC) as "ambo memberi makan ikan di kolam paman". There is a difference between the translation results from the random parallel corpus and the correct translation. This error can be corrected by optimized parallel corpus (OPC) to "dio ngasih makan ikan di kolam wak ".

Table 8. Examples of input sentences and output sentences of the translation results

No		Kalimat
1	Input	saya tidak mencuci baju di sungai
	MTS	ambo idak nyuci baju di sungai
	RPC	ambo idak mencuci baju di sungai
	OPC	ambo idak nyuci baju di sungai
2	Masukan	dia memberi makan ikan di kolam paman

	MTS	dio ngasih makan ikan di kolam wak
	RPC	ambo memberi makan ikan di kolam paman
	OPC	dio ngasih makan ikan di kolam wak
3	Masukan	jangan kamu berbicara seperti itu karena tidak baik di dengar orang lain
	MTS	jangan kau ngecek cak itu kareno idak baik di dengar orang lain
	RPC	jangan ambo berbicara seperti ambo karena idak baik di dengar orang lain
	OPC	jangan kau ngecek cak itu kareno idak baik di dengar orang lain
4	Input	cabe di pasar hari ini sangat mahal kata ibu saya
	MTS	cabe di pekan ari iko mahal nian kecek mak ambo
	RPC	cabe di pasar iko sanak iko mahal kata idak tau ambo
	OPC	cabe di pekan ari iko mahal nian kecek mak ambo

5. CONCLUSION

Parallel corpus optimization can increase the value of machine translation accuracy of Indonesian translators - Bengkulu Malay languages.

6. ACKNOWLEDGMENTS

Thank you to those who have helped in collecting the Indonesian language corpus data into Bengkulu Malay so that it can be collected to 3000 lines.

7. REFERENCES

[1] M. G. Asparilla, H. Sujaini, and R. D. Nyoto, "Perbaikan Kualitas Korpus untuk Meningkatkan Kualitas Mesin Penerjemah Statistik (Studi Kasus : Bahasa Indonesia – Jawa Krama)," vol. 1, no. 2, pp. 66–74, 2018.

[2] P. Permata and Z. Abidin, "Statistical Machine Translation Pada Bahasa Lampung Dialek Api Ke Bahasa Indonesia," *J. Media Inform. Budidarma*, vol. 4, no. 3, p. 519, 2020, doi: 10.30865/mib.v4i3.2116.

[3] L. Specia, *Statistical machine translation*, no. May 2012. 2012. doi: 10.4018/978-1-4666-2169-5.ch004.

[4] T. Apriani, H. Sujaini, and N. Safriadi, "Pengaruh kuantitas korpus terhadap akurasi mesin penerjemah statistik bahasa Bugis Wajo ke bahasa Indonesia," *JUSTIN (Jurnal Sist. dan Teknol. Informasi)*, vol. 1, no. 1, pp. 1–6, 2016.

[5] S. Mandira, H. Sujaini, and A. B. Putra, "Perbaikan Probabilitas Lexical Model Untuk Meningkatkan Akurasi Mesin Penerjemah Statistik," *J. Edukasi dan Penelit. Inform.*, vol. 2, no. 1, pp. 3–7, 2016, doi: 10.26418/jp.v2i1.13393.

[6] F. Rahutomo, R. A. Asmara, and D. K. P. Aji, "Computational analysis on rise and fall of Indonesian vocabulary during a period of time," *2018 6th Int. Conf. Inf. Commun. Technol. ICoICT 2018*, vol. 0, no. c, pp. 75–80, 2018, doi: 10.1109/ICoICT.2018.8528812.

[7] H. Ardhi, H. Sujaini, and A. B. Putra, "Analisis Penggabungan Korpus dari Hadits Nabi dan Alquran untuk Mesin Penerjemah Statistik," *J. Linguist. Komputasional*, vol. 1, no. 1, p. 31, 2018.

[8] R. Nugroho Aditya, T. Adji Bharata, and B. Hantono S, "Penerjemahan Bahasa Indonesia dan Bahasa Jawa Menggunakan Metode Statistik Berbasis Frasa," *Semin. Nas. Teknol. Inf. dan Komun.*, vol. 2015, no. Sentika, 2015.

[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. A. Suryani, D. H. Widyantoro, A. Purwarianti, and Y. Sudaryat, "Experiment on a phrase-based statistical machine translation using PoS Tag information for Sundanese into Indonesian," *2015 Int. Conf. Inf. Technol. Syst. Innov. ICITSI 2015 - Proc.*, 2016, doi: 10.1109/ICITSI.2015.7437678.

[11] D. Soyusiawaty and A. H. S. Jones, "Pemanfaatan Bahasa Alami Dalam Penelusuran Informasi Skripsi Melalui Digital Library," *Mob. Forensics*, vol. 2, no. 1, pp. 22–31, 2020, doi: 10.12928/mf.v2i1.2040.

[12] K. M. Shahih and A. Purwarianti, "Utterance disfluency handling in Indonesian-English machine translation," *4th IGNITE Conf. 2016 Int. Conf. Adv. Informatics Concepts, Theory Appl. ICAICTA 2016*, pp. 0–4, 2016, doi: 10.1109/ICAICTA.2016.7803104.

[13] M. Aadil and M. Asger, "An Overview of Statistical Machine Translation Tools," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 7, no. 7, p. 289, 2017, doi: 10.23956/ijarcsse/v7i7/0201.

[14] A. Wibawa, "Indonesian-to-Javanese Machine Translation," *Int. J. Innov. Manag. Technol.*, vol. 4, no. 4, pp. 451–454, 2013, doi: 10.7763/ijimt.2013.v4.440.

HASIL CEK_Dewi Soyusiawaty, Bella Okta Sari Miranda

ORIGINALITY REPORT

9%	6%	7%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	docplayer.net Internet Source	3%
2	Z Abidin, Permata, I Ahmad, Rusliyawati. "Effect of mono corpus quantity on statistical machine translation Indonesian – Lampung dialect of nyo", Journal of Physics: Conference Series, 2021 Publication	3%
3	e-journals.unmul.ac.id Internet Source	2%
4	Submitted to Tshwane University of Technology Student Paper	2%

Exclude quotes	On	Exclude matches	< 2%
Exclude bibliography	On		