pijmath

Pattimura International Journal of Mathematics (PIJMath)

Volume 02, issue 02, pages 53-62, November 2023
E-ISSN : 2830-6791

RESEARCH ARTICLE ◙ *OPEN ACCESS*

DOI: https://doi.org/10.30598/pijmathvol2iss2pp53-62

Application of Neural Machine Translation with Attention Mechanism for Translation of Indonesian to Seram Language (Geser)

Abdul Wahid Rukua[1], Yopi Andry Lesnussa[2], Dorteus Lodewyik Rahakbauw[3*], Berny Pebo Tomasouw[4]

[1,2,3,4] *Mathematic Program Study, Faculty of Mathematics and Natural Science, Universitas Pattimura*
Jln. Ir. M. Putuhenna, Ambon, 97233, Indonesia

Corresponding author's e-mail: [3] lodewyik@gmail.com*

ABSTRACT

Article History

Received : 23rd June 2023
Accepted : 5th September 2023
Published : 5th November 2023

Keywords

Neural Machine Translation (NMT);
Indonesia Language;
Seram Language (Geser).

The Seram language (Geser) is one of the regional languages in Kabupaten Seram Bagian Timur of Maluku Province which has been classified by the Language Office as an endangered language. This study uses the Neural Machine Translation (NMT) method in an effort to preserve the Seram (Geser) language. The NMT method has proven to be effective compared to SMT in overcoming the challenges of language translation by using the attention mechanism to improve translation accuracy. The data used in this study were obtained through interviews of 3538 parallel corpus, 255 Indonesian vocabularies and 269 Seram (Geser) vocabularies. The result showed that using 708 test data without Out-of Vocabulary (OOV) the BLUE Score was 0.90518992895191 or 90.518%.

[1] ***How to cite this article:***
Rukua A. W.,, Lesnussa Y. A., Rahakbauw D. L., & Tomasouw B. P., "APPLICATION OF NEURAL MACHINE TRANSLATION WITH ATTENTION MECHANISM FOR TRANSLATION OF INDONESIAN TO SERAM LANGUAGE (GESER)", *Pattimura Int. J. Math. (PIJMATH).*, vol. 02, iss. 02, pp. 53-62, November, 2023.

1. Introduction

Humans need communication to live and one of the communication tools is language [1]. Language is a form of crystallization of civilizational values and has an important role in directing the movement of civilization [2]. With the development of language studies, there is an increase in understanding and level of thought in a civilization [3]. Therefore, it is very important to learn a language in order to understand the culture, values and social background of the people who speak the language [4].

Cultural and linguistic diversity has long been one of Indonesia's wealth [5]. There is no exact information regarding the total number of languages spoken in the world, but it is estimated that about ten percent of them originate from Indonesia [6]. There are 726 languages in Indonesia [7], or according to the Language Development and Bookkeeping Agency of the Ministry of Education and Culture in 2019 there are 718 languages from 1340 ethnicities and tribes in Indonesia. This ethnic and cultural diversity creates linguistic phenomena such as bilingualism and language transformation until language extinction occurs [8].

The inability of languages in the eastern part of Indonesia to adapt to the dynamics of the times has made these languages vulnerable to extinction [9]. In the Maluku Province, Maluku Utara and Papua there are 11 out of 71 languages that have been declared extinct [10]. Meanwhile, according to the Maluku Provincial Language Office, it states that there are 22 languages in Maluku, one of which is the Seram language (Geser) [11].

Machine Translation is a technology that automatically converts from one language to another [12]. Recently, machine translation has achieved significant success [13]. Statistical Machine Learning (SMT) that has dominated machine learning research over the last few decades has been taken in its place by NMT (Neural Machine Translation) [14]. From a study of 50 million to 10 million paired sentences, NMT is superior to SMT [15].

NMT (Neural Machine Translation) is an end-to-end learning method for translating automatically [16], The main form of NMT consists of two main components: (1) The encoder calculates the c representation of each input sequence (2) Decoder that guesses word by word using information from c [17]. To make the encoder-decoder model more flexible, an attention mechanism is introduced [18]. From the NMT research using the attention mechanism for translating Lampung language into Indonesian, it shows a BLUE score of 51.96% with a total of 3,000 parallel corpus sentences and the best dimensional configuration [19].

Based on several studies and existing backgrounds, the researcher is interested in conducting research on how to apply Neural Machine Translation (NMT) for translating Seram (Slide) into Indonesian using the Attention mechanism.

2. Research Methods

2.1 Neural Machine Translation

The main form of NMT consists of two main components: (1) Encoder calculates the representation c of each input sequence (2) Decorder which guesses word by word using the information from c which is formulated as follows:

$$\log p(y|x) = \sum_{j=1}^{M} log\, p(y_j|y_{t<j}, c) \tag{1}$$

To model this equation, an RNN (Recurrent Neural Network) architecture can be used. The main components required in the BLSTM architecture are the same as NMT in general with a slight difference, namely: (1) Encoder (2) Global Attention (3) Local Attention (4) Decoder [20].

2.2 Research Subject

The subject of this research is a parallel corpus of Indonesian to Seram language (Geser) dialect spoken by people in Kabupaten Seram Bagian Timur (SBT), Provinsi Maluku.

2.3 Data Collection Techniques

The research data collection used interview method to obtain the data needed in the research. Researchers collected Geser language sentences and their translations by conducting interviews with 10 respondents (Geser language speakers).

2.4 The Tools Used

The tools used in this study are divided into several parts as follows:

a. NMT training platform: the platform used is Google Colab which is available for free and paid on the internet.
b. Hardware: the hardware used is 1 unit of 4.00 GB RAM computer with Intel ® Core ™ i3-6100 U Processor with CPU @ 2.30 Hz and built-in GPU from Google Colab, namely NVIDIA Tesla K80 which can only be used in 12 hours per day.
c. Deep Learning Framework: The framework used is Tensorflow.

2.5 Conceptual Framework

Figure 1 illustrates the stages of NMT research starting from the data collection stage to the conclusion stage.

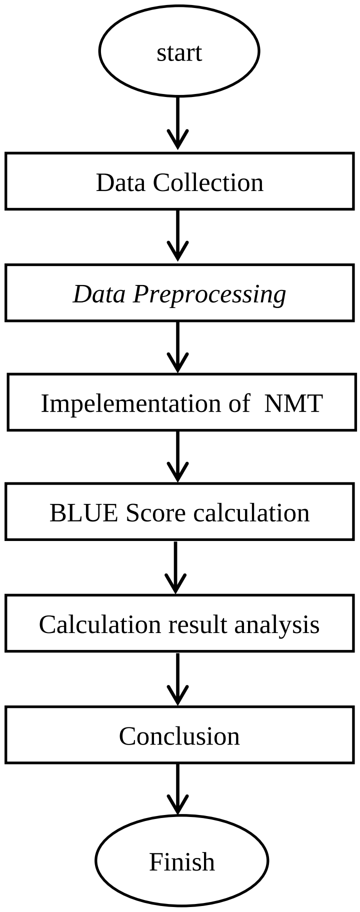


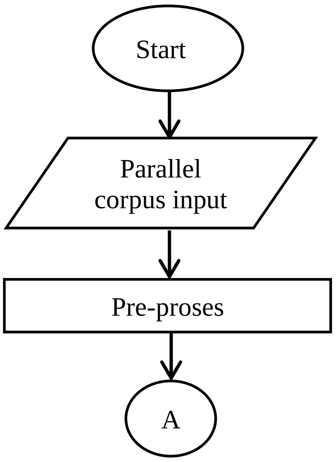
Figure 1. Conceptual Framework

2.6 Data Preprocessing

Before training, the dataset will be prepared in order to maximize the results of the training. This process consists of Data Sharing, Batching, Cleaning, Tokenazing and Text Vectorazation, Word Dictionary Creation, and Padding.

2.7 NMT System Design

The Neural Machine Translation (NMT) architecture used is Bidirectional Encoder-Decoder RNN for Geser Language to Indonesian translation. The process consists of parallel corpus input, pre-processing, model initialization, training, model evaluation, parameter tuning, best model and calculate BLEU Score.



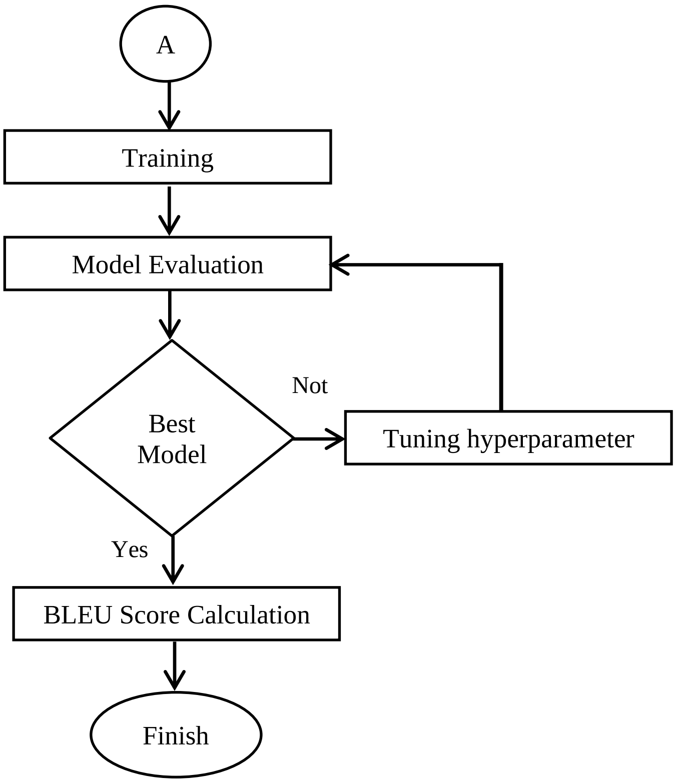


Figure 2. NMT System Design

3. Results And Discussion

The data obtained from the interviews amounted to 3538 parallel corpus, 255 Indonesian vocabularies and 269 Slang vocabularies which were used as the basis for testing the developed Neural Machine Translation (NMT) system.

3.1 Parallel Corpus Description

The parallel corpus contains pairs of original sentences to translated sentences consisting of several sentence types, namely: Verb Phrases and Verb Clauses. Here is an example of a parallel corpus of Indonesian – Geser language

Table 1. Example of Parallel Corpus Indonesia – Geser Language

No	Indonesia	Geser	Phrases / Clauses
1	Belajar Bahasa	Balajar Minak	Verb Phrase
2	Mengambil Pena	Na Kalam	Verb Phrase
3	Saya makan roti	Aku ka barot	Clause Verb
4	Ibu pergi ke pasar	Nina na tagi bua pasar	Clause Verb
5	Adik sedang belajar aljabar	Toi na balajar aljabar liwa	Clause Verb
	Dataset Total	**3538**	

Source : Primary data of interview results

The following is a visualization of the word length distribution in the Indonesian – Geser Language parallel corpus:

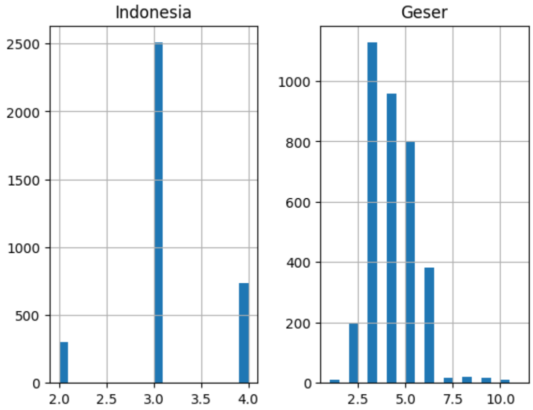


Figure 3. Sentence distribution on corpus Parallel

On **Figure 3**, Showing the difference in linguistic structure between Indonesian and Seram language (Geser) is very striking between these two languages. This is in accordance with what is shown in **Table 1** where the grammar, words, vocabulary, syntax, and sentence structure of these two languages are different.

3.2 Research Data Preparation

Before training there are several data preprocessing steps that must be done as follows:

1. Spliting Dataset

In this process the dataset is divided into two parts, namely for training and validation. This process is carried out using the library from Scikit-learn where the dataset is divided into 60% for training, 20% for validation and 20% for testing shown in **Figure 4** as follows:

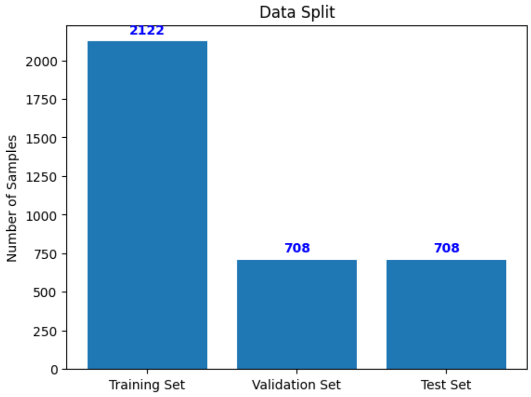


Figure 4. *Spliting Data* Result

2. Batching

Effective batching is 32 and 64. In this process, batching will be done with the Tensorflow Library. Here is the syntax:

```
BATCH_SIZE = 32
train_raw = tf.data.Dataset.from_tensor_slices((context_raw, target_raw)).batch
(BATCH_SIZE)
val_raw = tf.data.Dataset.from_tensor_slices((context_raw, target_raw)).batch(B
ATCH_SIZE)
```

3. Data Cleaning

In this process, the data is cleaned from errors such as typos, missing or duplicate data, and invalid or meaningless data. The data cleaning process is done using the following syntax:

```
def tf_lower_and_split_punct(text):
  text = tf_text.normalize_utf8(text, 'NFKD')
  text = tf.strings.lower(text)
  text = tf.strings.regex_replace(text, '[^ a-z0-9.?!,¿\\-+×÷]', '')
  text = tf.strings.regex_replace(text, '[.?!,¿]', r' \0 ')
  text = tf.strings.strip(text)
  text = tf.strings.join(['[START]', text, '[END]'], separator=' ')
  return text
```

4. Word Dictionary Creation

In this process the dataset is sorted as unique words or tokens of a sentence. The following is the word dictionary generation syntax:

```
context_vocab = np.array(context_text_processor.get_vocabulary())
tokens = context_vocab[example_tokens[0].numpy()]
' '.join(tokens)
```

5. Padding

In this process, datasets have different lengths. When padding is applied, a number of "padding" tokens will be added to the input sequence, so that all input sequences have the same length.

```
train_ds = train_raw.map(process_text, tf.data.AUTOTUNE)
val_ds = val_raw.map(process_text, tf.data.AUTOTUNE)
```

3.3 NMT Training

1. Training Simulation

In this study, several training simulations will be tried to evaluate the performance of the model. The following hyperparamater will be used to obtain the best model:

Table 2. Hyperparameter

No	Activation Function	Unit	Batch Size
1	RMSprop	128	32
2	Adagrad	256	
3	Adam	300	
4		512	

Table 3. Research Results

Activation Function	Unit	Loss	Masked Acc	Masked Loss	Val Loss	Val Masked Acc	Val Masked Loss
RMSprop	128	0.0456	0.9908	0.0456	0.1129	0.9797	0.1169
RMSprop	256	0.0650	0.9849	0.0650	0.1447	0.9708	0.1145
RMSprop	300	0.0481	0.9919	0.0481	0.1889	0.9587	0.1868
RMSprop	512	0.0955	0.9790	0.0955	0.2002	0.9529	0.2046
Adagrad	128	2.7273	0.4257	2.7273	27629	0.4409	2.7756
Adagrad	256	2.3587	0.4978	2.3587	2.4114	0.4894	2.4230
Adagrad	300	2.2671	0.5110	2.2671	2.3164	0.5005	2.3250
Adagrad	512	1.9104	0.5678	1.9104	1.9756	0.5493	1.9824
Adam	128	0.0158	0.9976	0.0158	0.0939	0.986	0.0912
Adam	256	0.0335	0.9937	0.0335	0.0947	0.9859	0.0927
Adam	300	0.0372	0.9935	0.0372	0.0981	0.9848	0.0952
Adam	512	0.0955	0.9790	0.0955	0.2002	0.9529	0.2046

On **Table 3**, shows that Adam's activation function with 128 units performs very well compared to others.

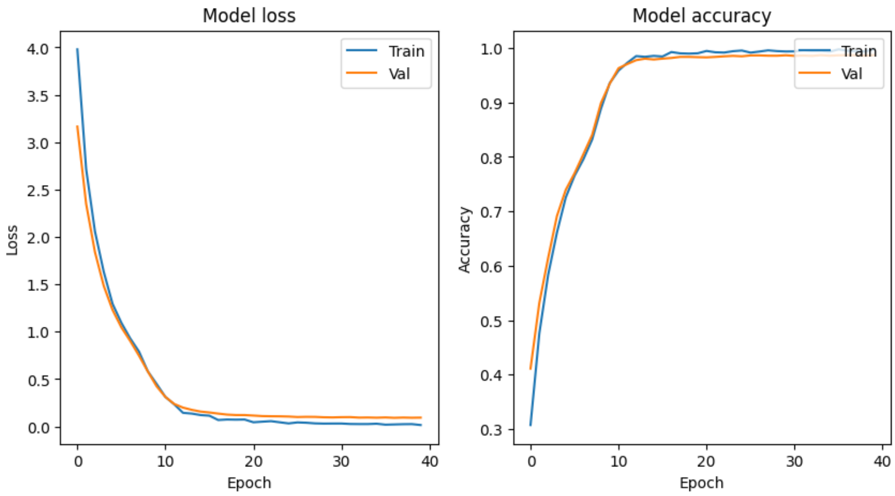


Figure 5. Accuracy and Loss Graph

On **Figure 5,** The left part shows the loss graph and the right part shows the accuracy graph of the model. Based on this information, it can be concluded that the model learned the patterns in the data well in the first few epochs, experienced significant improvement in the 10th epoch, and reached a relatively stable level of accuracy and loss at around epoch 40. The model also did not experience overfitting, which indicates its ability to generalize to new data.

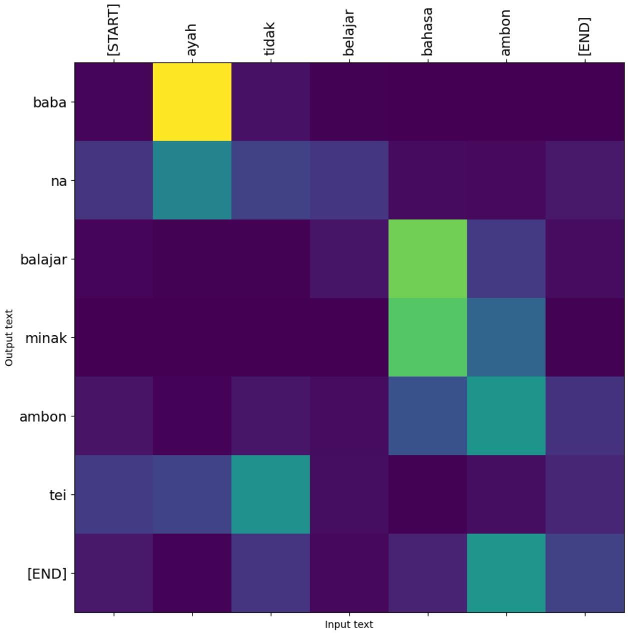


Figure 6. Attention Mechanism Chart

On **Figure 6** showed that the engine gave high attention scores for the words "ayah" and "bahasa", and this resulted in an excellent translation.

2. BLUE Score Testing

In the training simulation conducted, testing was carried out by trying to translate Indonesian sentences into Seram Language (Slide) machines without OOV (Output-out vocabulary). In the Blue Score test conducted using 708 test data, the BLUE Score value is 0.90518992895191 or 90%, this shows that the model has a very good accuracy rate. The following are the results of the machine and human translation comparison shown in **Table 4**, as follows:

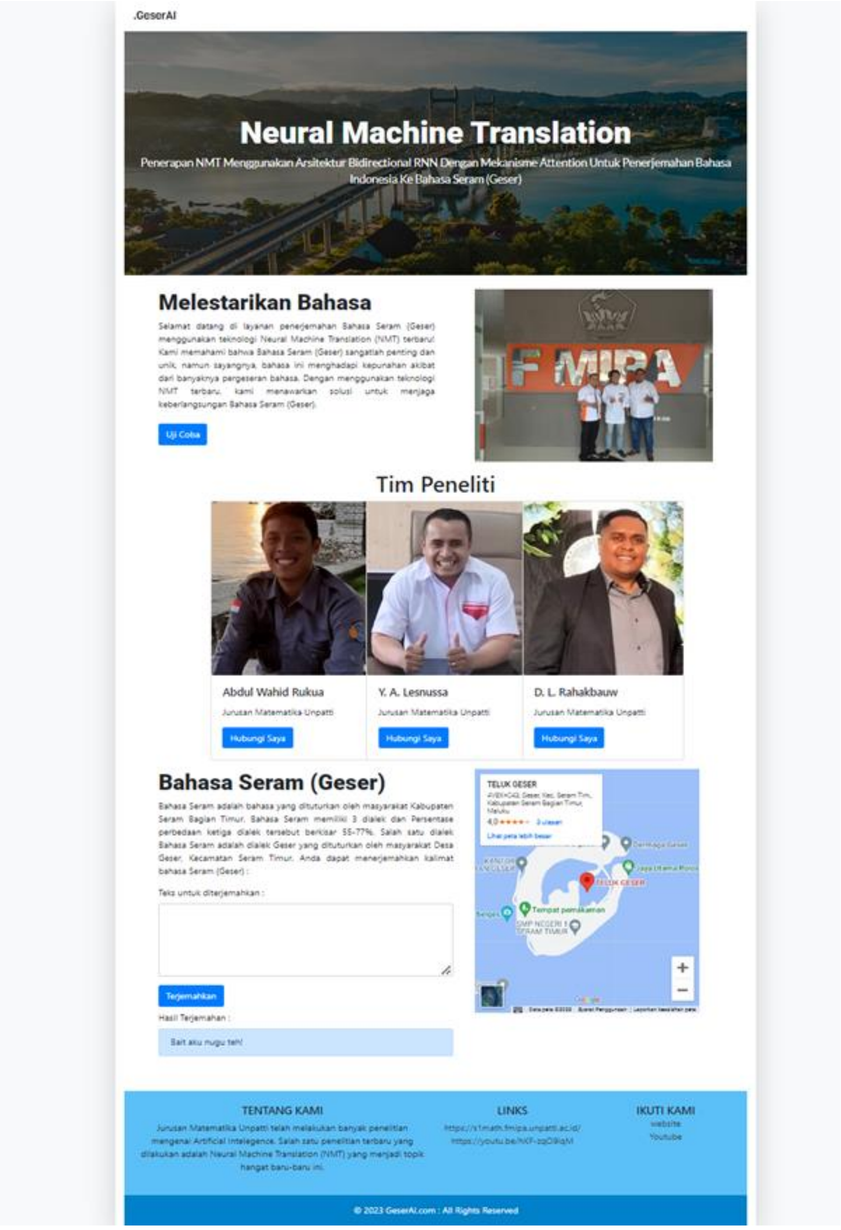
Table 4. Comparison of Machine and Human Translation

No	Indonesia	Machine Translation	Human Translation
1	Mengambil Mangkuk	Na basi kua	Na basi kua
2	Adik Belajar Bahasa Ambon	Toi na balajar manuk ambon	Toi na balajar manuk ambon
3	Aku Makan Mangga	Aku Ka Ayai	Aku Ka Ayai
4	Membeli Anggur	Fas Anggur a	Fas Anggur a

3. Application of Website-based NMT

Interface design plays an important role in improving the effectiveness of Neural Machine Translation (NMT) models. The use of an appropriate interface can enhance the user's ability to interact with the model more effectively, as well as improve the performance of the model. As a result, much research has been done to develop an optimal interface design for NMT models.

The development of this website uses the best model from NMT training results stored in tensorflow tf-lite and then deployed using the Flask framework The following is an image of the NMT Interface Design for Indonesian to Seram language translation (Geser).



.GeserAI

Neural Machine Translation

Penerapan NMT Menggunakan Arsitektur Bidirectional RNN Dengan Mekanisme Attention Untuk Penerjemahan Bahasa Indonesia Ke Bahasa Seram (Geser)

Melestarikan Bahasa

Selamat datang di layanan penerjemahan Bahasa Seram (Geser) menggunakan teknologi Neural Machine Translation (NMT) terbaru! Kami memahami bahwa Bahasa Seram (Geser) sangatlah penting dan unik, namun sayangnya, bahasa ini menghadapi kepunahan akibat dari banyaknya pergeseran bahasa. Dengan menggunakan teknologi NMT terbaru, kami menawarkan solusi untuk menjaga keberlangsungan Bahasa Seram (Geser).

Tim Peneliti

Abdul Wahid Rukua
Junusan Matematika Unpatti

Y. A. Lesnussa
Junusan Matematika Unpatti

D. L. Rahakbauw
Junusan Matematika Unpatti

Bahasa Seram (Geser)

Bahasa Seram adalah bahasa yang dituturkan oleh masyarakat Kabupaten Seram Bagian Timur. Bahasa Seram memiliki 3 dialek dan Persentase perbedaan ketiga dialek tersebut berkisar 55-77%. Salah satu dialek Bahasa Seram adalah dialek Geser yang dituturkan oleh masyarakat Desa Geser, Kecamatan Seram Timur. Anda dapat menerjemahkan kalimat bahasa Seram (Geser) :

Teks untuk diterjemahkan :

Hasil Terjemahan :

Bait aku nugu teh!

TENTANG KAMI

Junusan Matematika Unpatti telah melakukan banyak penelitian mengenai Artificial Intelegence. Salah satu penelitian terbaru yang dilakukan adalah Neural Machine Translation (NMT) yang menjadi topik hangat baru-baru ini.

LINKS

https://s1math.fmipa.unpatti.ac.id/
https://youtu.be/hKF-zqO8lqM

IKUTI KAMI

website
Youtube

© 2023 GeserAI.com : All Rights Reserved

Figure 7. Interface NMT Design

4. Conclusions

From the results of research conducted using 708 test data, the BLUE Score value is 0.90518992895191 or 90% for test data without Out-of Vocabulary (OOV).

References

[1] O. Mailani, I. Nuraeni, S. A. Syakila, and J. Lazuardi, “Bahasa Sebagai Alat Komunikasi Dalam Kehidupan Manusia,” [Language as a Communication Tool in Human Life] *Kampret Journal*, vol. 1, no. 1, pp. 1–10, Jan. 2022, doi: 10.35335/kampret.v1i1.8.

[2] I. Siregar, “International Journal of Cultural and Religious Studies (IJCRS) The Existence of Culture in its Relevance to the Dynamics of Globalization: Bahasa Indonesia Case Study,” 2021, doi: 10.32996/ijcrs.

[3] B. Ashcroft, “Language and Race,” *Social Identities*, vol. 7, no. 3, pp. 311–328, Sep. 2001, doi: 10.1080/13504630120087190.

[4] C. Kramsch, “Language and Culture,” *AILA Review*, vol. 27, pp. 30–55, Dec. 2014, doi: 10.1075/aila.27.02kra.

[5] Ramot Peter and Masda Surti Simatupang, “Keberagaman Bahasa Dan Budaya Sebagai Kekayaan Bangsa Indonesia,” *Dialektika: Jurnal Bahasa, Sastra, dan Budaya*, pp. 96–105, 2022.

[6] Hein Steinhauer, “The Indonesian language situation and linguistics: Prospects and possibilities,” *Bijdragen tot de taal-, land- en volkenkunde*, pp. 755–784, 1994.

[7] D. Crystal, *Language Death*. Cambridge University Press, 2002. doi: 10.1017/CBO9781139871549.

[8] F. H. Tondo, “Kepunahan Bahasa-Bahasa Daerah: Faktor Penyebab Dan Implikasi Etnolinguistis 1,” [The Extinction Of Regional Languages: Causal Factors And Ethnolinguistic Implications] 2009.

[9] Tim WowKeren, “Pakar Bahasa Menilai Bahwa Bahasa Bisa Punah Lantaran Mulai Ditinggalkan Oleh Penuturnya. Hal Itu Disebabkan Oleh Faktor Tertentu, Misalnya Ketidakmampuan Bahasa Dalam Menyesuaikan Diri.,” [Language experts believe that languages can become extinct because they are abandoned by their speakers. This is caused by certain factors, for example the inability of language to adjust itself] Feb. 20, 2020. https://www.wowkeren.com/berita/tampil/00298629.html (accessed Jun. 08, 2023).

[10] S. Zulfiqar Bin-Tahir, H. Hanapi, N. Mufidah, A. Rahman, and V. U. Tuharea, “Revitalizing The Maluku Local Language In Multilingual Learning Model, “*International Journal Of Scientific & Technology Research*, vol. 8, no. 10, 2019, [Online]. Available: www.ijstr.org

[11] RERE KHAIRIYAH, “Hasil Penelitian, 7 Bahasa Daerah di Maluku Punah, 22 Terancam,” [Research Results, Seven Regional Languages in Maluku Extinct, 22 Threatened] Jul. 29, 2017. https://nasional.tempo.co/read/904371/hasil-penelitian-7-bahasa-daerah-di-maluku-punah-22-terancam (accessed May 31, 2023).

[12] Dorothy Kenny, *Machine translation*, 3rd Edition. Routledge, 2019.

[13] M. Mager, E. Mager, K. Kann, and N. T. Vu, “Ethical Considerations for Machine Translation of Indigenous Languages: Giving a Voice to the Speakers,” May 2023.

[14] F. Stahlberg, “Neural Machine Translation: A Review,” *Journal of Artificial Intelligence Research*, vol. 69, pp. 343–418, Oct. 2020, doi: 10.1613/jair.1.12007.

[15] S. Kinoshita, T. Oshio, and T. Mitsuhashi, “Comparison of SMT and NMT trained with large Patent Corpora: Japio at WAT2017,” 2017. [Online]. Available: https://bitbucket.org/eunjeon/mecab-ko/

[16] Y. Wu *et al.*, “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation,” Sep. 2016.

[17] E. A. Gemechu and G. R. Kanagachidambaresan, “Machine Learning Approach to English-Afaan Oromo Text-Text Translation: Using Attention based Neural Machine Translation,” in *2021 4th International Conference on Computing and Communications Technologies (ICCCT)*, IEEE, Dec. 2021, pp. 80–85. doi: 10.1109/ICCCT53315.2021.9711807.

[18] S. Feng, S. Liu, M. Li, and M. Zhou, “Implicit Distortion and Fertility Models for Attention-based Encoder-Decoder NMT Model,” Jan. 2016.

[19] Z. Abidin, A. Sucipto, A. Budiman, F. Teknik, and D. I. Komputer, “Penerjemahan Kalimat Bahasa Lampung-Indonesia Dengan Pendekatan Neural Machine Translation Berbasis Attention Translation Of Sentence Lampung-Indonesian Languages With Neural Machine Translation Attention Based Approach”, [Translation Of Sentence Lampung-Indonesian Languages With Neural Machine Translation Attention Based Approach Translation Of Sentence Lampung-Indonesian Languages With Neural Machine Translation Attention Based Approach]. Available: https://translate.google.com.

[20] M.-T. Luong, H. Pham, and C. D. Manning, “Effective Approaches to Attention-based Neural Machine Translation,” Aug. 2015.