MATICS: Jurnal Ilmu Komputer dan Teknologi Informasi ISSN : 1978-161X(p); 2477-2550(e)
(*Journal of Computer Science and Information Technology)* *Volume* 14, No. 1(2022), pp 13-21
Received: January25th 2022; **Accepted**: March 9th 2022; **Avalaible Online**: June15th 2022

Phrase Based and Neural Network Translation for Text Transliteration from Arabic to Indonesian

Alvian Burhanuddin, Ahmad Latif Qosim, Rizqi Amaliya, Muhammad Faisal

***Abstract*— Transliteration is one solution to overcome the inability to read and write Arabic in Indonesia. However, this transliteration has many different versions in reality. The many different transliterated versions make it difficult for people to understand and pronounce Arabic sentences. In this study, the researcher presents a comparison of the data mining approach to transliteration. Phrase based machine translation (PBSMT) and attention based are the methods we compare in this study. In doing this transliteration, PBSMT has a better value.**

***Index Terms*— Transliteration; data mining; Indonesian language; phrase-based machine translation; neuro machine translation; attention-based; statistical machine based;.**

I. INTRODUCTION

Based on the census conducted by the Central Statistics Agency, almost 87 per cent of Indonesia's population embraces Islam[1]. Almost all of the religious activities of Muslims use Arabic. Ironically, a survey conducted by IIQ noted that 65% of Muslims are blind to the Koran[12]. Depart from that background; one solution is to transliterate Arabic into Indonesian.

The existence of this transliteration process is something that the academician must accelerate. Given that this need is an intermediary in carrying out worship in Islam itself, it Includes dhikr, wirid, and other religious activities. However, when there is a transliteration, it should not learn independently without someone directing it. Because there are some parts of the transliteration that cannot match the pronunciation or the original sound of the Arabic reading. Al-Faruqi[2] noted that the primary purpose of transliteration of Arabic is not transliteration itself but makes it easier to learn it. Furthermore, if the transliteration is the Koran, then the transliteration cannot be said to be the Koran. In addition, when looking at the letter Yusuf verse 2, the meaning is "*Indeed, we have sent down the Qur'an in Arabic, I hope you think about it.*"

In its development, the main problem of transliterating Arabic manuals into Indonesian is a large number of references[3], [4]. This transliteration causes inconsistencies between transliterations with each other. For example, in writing the letter خ, which if transliterated, it will have several writings. Including kh, h, and cho. As a result, when the reader moves from one transliteration to another. That will not be easy to understand the spelling.

Several approaches to transliteration from one language to another. Virga[25] transliterated the name from English into Chinese script. The research resulted that the machine translation method could improve the text transliteration process.

Seeing the problem's urgency and the lack of transliteration research, the researcher wants to transliterate Arabic into Indonesian using a data mining approach. As a consideration, the researcher also made a comparison of two machine learning methods. It aims to find the best model for transliteration. Thus, it is expected that there will be benefits for academics in the form of a reference for academics in developing machine-based Indonesian language models. In this research, there are also benefits for the general public in ease of understanding and reading texts because all letters use a uniform language.

A. Transliteration

Transliteration is defined as the copying or replacement of letters from one alphabet to another[5]. The ministerial decree[6] translates transliteration as a transfer of letters from one alphabet to the others and its equipment. So it can be concluded that transliteration is a way to change letters from one alphabet to another with matters related to the letter.

Manuscript received March 22, 2007. (Write the date on which you submitted your paper for review.) This work was supported in part by Informatics Engineering Department of Maulana Malik Ibrahim Islamic State University..

Alvian Burhanuddin is with the Informatic Engineering Departement of Maulana Malik Ibrahim Islamic State University , Malang, Indonesia (alvianthelfarqy@gmail.com)

Ahmad Latif Qosim is with the Informatic Engineering Departement of Maulana Malik Ibrahim Islamic State University , Malang, Indonesia (19841007@student.uin-malang.ac.id)

Rizqi Amaliya is with the Informatic Engineering Departement of Maulana Malik Ibrahim Islamic State University , Malang, Indonesia (19841008@student.uin-malang.ac.id)

Some formulas for transliteration of Arabic into Indonesian have also been explained in the minister's decision document [6] Including:

1. Consonants
2. Vowels
3. Maddah
4. Ta'marbutah
5. Syaddah
6. Harakah
7. Hamzah
8. Word Writing
9. Capital Letters
10. Tajwid

The main table as the basis for transliterating Arabic to Indonesian can also be founded in table 1.

Table 1. Transliteration Table

Arabic letters	Name	Indonesian letters	Name
ا	alif	Not denoted	Not denoted
ب	ba	B	be
ت	ta	T	Te
ث	sa	s	Es
ج	jim	J	Je
ح	ha	H	Ha
خ	kha	Kh	Ka dan ha
د	dal	D	De
ذ	zal	Dz	Dzet
ر	ra	R	Er
ز	zai	Z	Z
س	sin	S	Es
ش	syin	Sy	Es dan ye
ص	shad	Sh	Es dan ha
ض	dhad	Dh	De dan ha
ط	tha	Th	Te dan ha
ظ	zha	Zet	Zet
ع	ain	‘	upside-down comma above
غ	gain	G	Ge
ف	fa	F	Ef
ق	qaf	Q	Qi
ك	kaf	K	Ka
ل	Lam	L	El
م	mim	M	Em
ن	nun	N	En
و	wau	W	We
ه	ha	H	H
ء	hamzah	`	apostrophe
ي	ya	Y	ye

Further research found that many other transliterations emerged along with the adaptation of society. Another finding of transliteration was disclosed by Ahmad[3], although transliteration from Arabic to Indonesian can help people who cannot read Arabic. Mainly in the matter of Hajj rituals. The circulation of various transliteration models has resulted in irregularities between one transliteration guideline and other transliteration guidelines. This circulation makes it difficult for people to read transliteration. Several suggestions for improvement in transliteration have been given, namely providing transliteration that accommodates the original sound instead of the character. However, such suggestions require a review of the current transliteration regulations.

II. Related Research

Guellil[7] transliterates from arabizi to Arabic using machine translation. Arabizi is the Latin language which is a transliteration of Arabic using the Algerian dialect. This study uses datasets from several social media such as Facebook, Twitter, and Youtube. Statistical machine translation and neural machine translation produce 73% and 75% accuracy, respectively, on internal data. Meanwhile, on external data, it was found that the accuracy was only 45%. However, this study increased accuracy by up to 2%.

Ameur[8] also conducted research using transliteration for Arabic into English. This study uses neuro machine translation with the use of RNN. In this study, some comparisons on the machine translation algorithm. The result is that the bidirectional attention-based encoder-decoder gets the highest score compared to other methods. For transliteration from English to Arabic, the result is a word error rate of up to 5.40 and a character error rate of 0.95. Meanwhile, for Arabic to English, the word error rate is 65.16, and the character error rate is 16.35. What is quite interesting from this research is that phrase-based machine translation gets the second-best value compared to more modern neural network algorithms.

Masmoudi[9] uses syntactical rules to create datasets. From the dataset, then transliteration using CRF was performed. The research, conducted on the Tunisian dialect, resulted in a word error rate score of 9.80 and a character error rate of 10.47. Compared to several previous methods, this hybrid approach can produce many datasets in a research initialization. So, even though the research has a relatively low score, this research is one of the best studies in the unsupervised learning process.

III. Methodology

The research procedure for conducting empirical tests can be found in Figure 1. The diagram starts from the dataset collection stage to the prediction process.

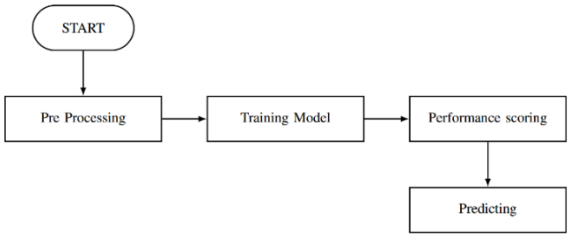

Figure 1. Research procedure diagram

A. Dataset

The transliterated dataset into Indonesian is one of the datasets that are rarely found. To deal with this situation, we use daily conversation sentences. The dataset in this study consists of 1000 sentences, with approximately 5000 words. To get the best accuracy, we manually label and check the datasets one by one.

B. Preprocessing

The diagram begins with the preprocessing process. In the preprocessing stage, the data cleaning stage is carried out. This cleaning phase includes omitting unnecessary parts, such as Arabic meanings, unnecessary quotes such as "(",")", and numbers. Instead, the researcher only took the Arabic text and its transliteration.

The absence transliteration dataset from Arabic to Indonesian adds to one of our tasks to build it. We started with manually checking the Arabic sentence data and transliteration. This check includes the existence of harakat and the suitability of the harakat. Then proceed with normalization.

In the normalization section, the researcher checked the possibility of paragraphs in the Arabic text data. When found, then the sentence is separated according to the period. If no punctuation is available, then they are separated by meaning. To maintain the credibility of the data, we do this part manually.

C. Model Training

After we get cleaned the transliteration dataset, Arabic text and transliteration text are formed into tokens. This token is used in the training model process. To get the best results, the researchers compared three machine translation methods using the dictionary method, Phrase-based Method, and Neural network. Then, by the researcher, each data was trained with the model.

1. Machine translation

Lopez[10] states that *Machine translation* can be defined as the automatic translation from one language to another by a computer. Machine learning methods are used to achieve the automation of the translation. First, a series of machine learning algorithms are applied to study the translated text dataset or corpus. The engine can then estimate other text data to be translated based on the value of the algorithm.

Technically, the main task of machine translation is to change the token set from the source language with vocabulary V_s. into the target language token set with vocabulary V_t. Assuming the token in the sequence is a word, while the sequence is a sentence. Thus, there are three stages in implementing machine translation:

1. Stages to describe how the language change will be carried out. Several algorithms for converting the original language into another language have existed in previous studies. Among them are word alignment[11], Phrase-based Machine Translation[11], and Neuro Machine Translation[12]
2. Parameterization. This process is a process for evaluating each data. This assessment is inseparable from machine learning rules, which uses a series of statistical models in its measurement.
3. Decoding. This process is related to the input provided by the user. For example, when the user inputs a sentence, the computer will search the highest value according to the data in the model. Accuracy and speed are general discussions at this stage.

The following section will explain the discussion related to these steps, especially in the algorithm section. By the researcher's description in the problem definition section, the section we discuss is the algorithm for translation. In the parameterization and decoding section, we use BLEU assessment and greedy decoding.

2. Phrase based machine translation

The Phrase-based Machine Translation model is based on noisy channels. So it can be written in the equation:

$$t_{best} = argmax_t\{Pr(t|s)\} \qquad (1)$$

Koehn[13] reformulated equation 2 using the Bayes rule. Where is the equation:

$$argmax_t p(t|s) = argmax_t p(t|s)p(s) \qquad (2)$$

from these equations, the model can be calculated separately, namely the translational probability of the model $p(t|s)$ and the probability of the source language model itself $p(t)$.

The equation is used to translate or change the word sequence in the original sentence to the translation on the target sentence. This equation is used to find the best value of the word candidates in the sentence. The best word candidates in question resulted from the translation process, in some cases experiencing a change in order. This change is influenced by the probability value between word sequences from the actual data used as a training model. Application of these naive rules is through the moses[11] framework, which is also used in this research.

There are at least three approaches to extracting phrases in sentences[13]. Namely by using phrases from Word Alignment, phrases from syntactic rules, and phrases from phrase alignments or commonly referred to as n-grams.

3. Word alignment

Word Alignment was first coined by Brown[14]. This idea writes simply that word pairs describe the translation of one sentence into another. In a translation, the order of the words is generalized. For example, in the translation of French to English, (*Le program an étémis en application | And the(1) program(2) has(3) been(4) implemented(5,6,7)*). The translation describes all words from English adapted to French. These French words are linked into English regardless of their arrangement. Instead, use the equivalent words in the dictionary.

In his subsequent research[15], Brown found that some words might not be appropriate due to differences in the structure of each language. To overcome this, Brown conducted a probability measurement on the similarity of words that often appear. This study found that the likelihood of words gave good results in the translation dataset from English to French.

As research continues to develop, several improvements to word alignment continue to be made. For example, Li[16] found that word alignment outperformed other

algorithms on specific datasets. This discovery is demonstrated by experimentation by applying the Prediction Difference method using a deterministic[17] and sampling[18] approach.

4. Neuro Machine translation

Neuro Machine Translation is one of the new methods in the development of translation technology. This method uses a series of sentences as input in the training process. Furthermore, the input is used as a reference for the calculation of the output prediction. In its design, NMT consists of 2 main parts, Encoder, which performs calculation of word value representation for each input. Moreover, the decoder, which finds the word pairs present on the input by performing probability calculations. In performing these probability calculations, the RNN architecture used by kalchbrenner[19] uses the equation.

$$logp(y|x) = \sum_{j=1}^{m} logp(y_j|x_{<j}, s) \quad (3)$$

As for the encoder, this research uses a convolutional neural network. On the other hand, bahdanau[20] uses the Gate Recurrent unit (GRU) approach, a development that differs from the general RNN method in its encoder and decoder. One of these uses is based on the effectiveness of the GRU, which is better than other traditional methods[21].

In conducting training with the attention model, the researcher encodes the input sentences using GRU. The results of the GRU encoding in the form of vectors and hidden states will then be calculated to get a context vector that has the same equation

$$c_t = \sum_s \alpha_{ts} \bar{h}_s \quad (4)$$

In this equation, alpha is the attention weight of a word in the sentence sequence. This alpha value is obtained from the equation:

$$\alpha_{ts} = \frac{exp(score(h_t, \bar{h}_s))}{\sum_{s'=1}^{S} exp(score(h_t, \bar{h}_s))} \quad (5)$$

$score(h_t, \bar{h}_s)$ calculates the last hidden state decoder value with the value of each hidden state encoder in each token. The equation for calculating this value is written with the information below this article.

$$score(h_t, \bar{h}_s) = v_\alpha^T tanh(W_a[h_t; \bar{h}_s]) \quad (6)$$

From the calculated value, the value of the context vector is calculated based on the overall mean value of hidden states.

The epoch learning in this research was weight update from one training iteration. As the result from all epoch on this training is model. The model generated from the training process is then validated and tested. This process is implemented to find out the best results from using the three methods above. In selecting value variables, we use the BLEU[22] metric. The BLEU metric compares the n-grams of the candidate with the reference. The BLEU value is obtained from the number of words in the target match with the reference.

IV. Discussion

In this section, the researcher describes the experimental results according to the method described in the previous section. So it can be seen the actual accuracy of machine translation in transliteration using Phrase-based machine learning models and neural networks.

Before training on the model, we aligned the dataset of sentences to be transliterated with transliterated sentences in Indonesian. In this alignment, we use MGIZA[23]. From the alignment, a sentence is separated by word. The results of this pair of words will be used as a training model to determine the weight value. Koehn[13] wrote that three things influence weight on the word. First, the heuristic suitability between word phrases in the source and target, the number of phrases in a sentence, and the word's suitability with the sentence's syntax.

Based on this, we conducted training using the 1-gram, 2-gram, and 3-gram arrangement in PBSMT. From the 3 data training methods, we found that 3-grams generated the best model with a BLEU value of up to 24.

The results obtained from the comparison of these methods are shown in the table 2. In the table, the researcher provides examples of testing data consisting of ground truth and the transliteration results of each model.

Table 2. Experiment Table

Arabic text	من بين ايديهم
ground truth	mim bayna aydihim
Statistical based model	mim bayna aydihim .
attention based model	mim bayna razaqna

Using the same dataset, we conducted training on an attention-based neural network-based model with Bahdanau's approach. We made several changes to the epochs and batches used in this training process, consisting of 8, 16, and 24. The best model was generated by train using 16 epochs. Attention-based in this study had the lowest bleu value compared to PBSMT. To find a bright spot in this condition, the researcher conducted manual observations on the model. The decrease in the bleu value was caused by the many changes in the word order in the sentence. This drop is because attention-based does not transliterate according to the words in the sentence.

From the results of this experiment, we found that the model with the highest BLEU score in this study was the approach with Phrase-based machine transliteration with a value of 24. However, in the experiments we conducted using this model, some words experienced changes from their fundamental values. So, one way to increase from this value is to be able to use more datasets. It is hoped that there will be an increase in the accuracy obtained.

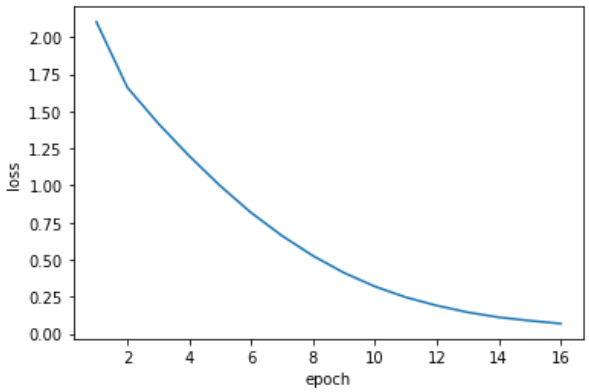

Figure 2. Loss graph uses 16 epoch

V. Conclusion

In this paper, researchers have explored machine learning in the transliteration process from Arabic to Indonesian. Researchers compare several machine learning methods that have existed before and have been commonly used in text mining processing. The models are neural networks with attention-based and statistical machine-based. Of the two models, the best results are obtained on the use of statistical-based. This research cannot be separated from its shortcomings, for example, in the part of the wording that is not following the ground truth. In addition, it is necessary to revisit the use of vowels, which is very influential on the transliteration process in Arabic into Indonesian and other languages. The development of this research, can be done by adding to the dataset used. In addition, fine-tuning also needs to be done using pre-existing datasets, such as transliteration for Malay.

References

[1] S. Indonesia, "badan pusat statistik," *BPS-Stat. Indones.*, 2018.
[2] I. R. Al-Faruqi, *Toward Islamic English*. International Institute of Islamic Thought (IIIT), 1986.
[3] N. F. Ahmad, "Problematika Transliterasi Aksara Arab-Latin: Studi Kasus Buku Panduan Manasik Haji dan Umrah," *Nusa J. Ilmu Bhs. Dan Sastra*, vol. 12, no. 1, pp. 126–136, 2017.
[4] M. Musadad, "Alquran transliterasi latin dan problematikanya dalam ma-syarakat muslim denpasar," *SUHUF*, vol. 10, no. 1, pp. 193–209, 2017.
[5] T. R. K. B. Indonesia, "Kamus Bahasa Indonesia," *Jkt. Pus. Bhs. Dep. Pendidik. Nas.*, vol. 725, 2008.
[6] K. B. M. Agama, M. Pendidikan, and K. R. Indonesia, "Nomor 158 Tahun 1987 dan Nomor 0543 b." Departemen Agama, Jakarta, 1987.
[7] I. Guellil, F. Azouaou, M. Abbas, and S. Fatiha, "Arabizi transliteration of Algerian Arabic dialect into modern standard Arabic," 2017.
[8] M. S. H. Ameur, F. Meziane, and A. Guessoum, "Arabic machine transliteration using an attention-based encoder-decoder model," *Procedia Comput. Sci.*, vol. 117, pp. 287–297, 2017.
[9] A. Masmoudi, M. E. Khmekhem, M. Khrouf, and L. H. Belguith, "Transliteration of arabizi into arabic script for tunisian dialect," *ACM Trans. Asian Low-Resour. Lang. Inf. Process. TALLIP*, vol. 19, no. 2, pp. 1–21, 2019.
[10] A. Lopez, "Statistical machine translation," *ACM Comput. Surv. CSUR*, vol. 40, no. 3, pp. 1–49, 2008.
[11] P. Koehn *et al.*, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, 2007, pp. 177–180.
[12] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, "Opennmt: Open-source toolkit for neural machine translation," *Prepr. ArXiv170102810*, 2017.
[13] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST, 2003.
[14] P. F. Brown *et al.*, "A statistical approach to machine translation," *Comput. Linguist.*, vol. 16, no. 2, pp. 79–85, 1990.
[15] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Comput. Linguist.*, vol. 19, no. 2, pp. 263–311, 1993.
[16] X. Li, G. Li, L. Liu, M. Meng, and S. Shi, "On the word alignment from neural machine translation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 1293–1303.
[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[18] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," *arXiv:1702.04595*, 2017.
[19] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1700–1709.
[20] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *ArXiv Prepr. ArXiv14090473*, 2014.
[21] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *ArXiv Prepr. ArXiv14123555*, 2014.
[22] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
[23] Q. Gao and S. Vogel, "Parallel implementations of word alignment tool," in *Software engineering, testing, and quality assurance for natural language processing*, 2008, pp. 49–57.
[24] Virga, Paola and Khudanpur, Sanjeev. "*Transliteration of proper names in cross-lingual information retrieval*" in *Transliteration of proper names in cross-lingual information retrieval*, 2003.