Malay Manuscripts Transliteration Using Statistical Machine Translation (SMT)

Sitti Munirah Abdul Razak, Muhamad Sadry Abu Seman, Wan Ali @ Wan Yusoff Wan Mamat, Noor Hasrul Nizan Mohammad Noor
Department of Information System
Kulliyyah of Information and Communication Technology
International Islamic University Malaysia
Gombak, 53100, Selangor
sittimunirahabdulrazak@gmail.com

***Abstract*—Natural Language Processing (NLP) is a vital field of artificial intelligence that automates the study of human language. However for Malay manuscripts (MM) written in old jawi, its exposure on such field is limited. Besides, most of the studies related to MM studies and NLP were focused on rule based or rule based machine transliteration (RBMT). Hence the objective of this study is to propose a statistical approach for old jawi to modern jawi transliteration of Malay manuscript contents using Phrase Based Statistical Machine Translation (PBSMT) as its model. In order to achieve such purpose, quality score of Word Error Rate (WER) was computed on the transliteration output. Besides, the issues formerly encountered by rule based approach such as vocals limitation and homograph, reduplication, letters error and combination of multiple words were observed in the implementation. Moreover, this paper utilized exploratory approach as its research strategy and mixed method as its research method. The data for the analysis were extracted from a MM titled *Bidāyat al-Mubtadī bi-Faḍlillāh al-Muhdī*. Quality score of WER was computed for the evaluation of SMT output. Afterwards, related issues were identified and assessed. The research found that quality score of PBSMT for old jawi to modern jawi transliteration was high in terms of WER, however the issues of rule based were generally addressed by PBSMT except homograph. The research is however limited to the approach of SMT that solely focused on PBSMT as its model. Moreover, the corpus size was limited to one manuscript while SMT relies on corpus size. Nevertheless the research contributes to the wider coverage on Malay language as one of the under resource languages in NLP, in form of old and modern jawi. Besides, to the best of the researcher's knowledge, it is also the first to apply SMT (PBSMT) approach on old jawi transliteration. Most importantly, the study is to contribute on MM's studies.**

Keywords—jawi, Malay manuscript, old jawi, Statistical Machine Translation, Phrase Based Statistical Machine Translation, Natural Language Processing

I. INTRODUCTION

In Malaysia, Malay Manuscript (MM) is among the highly valued heritages of the Malays. Yet, study of MM is limited and inefficient in the context of information technology, especially on text mining and artificial intelligence that have grown popular due to its tremendous advantages on both academic and industrial area. MMs are highly relevant to these fields due to its rich contents on various fields including of historical and religious knowledge that have not been fully discovered and continuously studied on. Unfortunately, MM's studies are still inadequate on the initial task of Natural Language Processing (NLP), which is data extraction. As MMs were written in old jawi which is one of the under resource languages, transliteration is required as part of the extraction process. One of the significances of manuscript transliteration is to convert the text into a more readable and comprehensible format, such as modern jawi or roman form [1]. Transliteration is similar to translation, except that translation converts a source language into another, while transliteration is monolingual. It instead converts one's writing system into another form yet similar in phonetic or sound [2][3]. Significantly instead of hiring professionals that consume time and cost, implementing NLP should automate the extraction of MMs written in old jawi text. As a result the study of MMs shall be more efficient and effective, as researchers will be able to process the content for automated and extensive learning.

There are two main approaches of machine learning for translations: statistical based (SMT) and rule based (RBMT) [4]. SMT can be described as a translation process inferred based on the probability statistics generated from the corpora whereas RBMT can be defined as a conversion process based on the sets of language rules defined by experts [4]. Contrary to translation studies on other languages that were more focused on SMT, transliteration studies of old jawi were more focused on rule based approach or RBMT. Significantly, RBMT is highly crucial for old jawi transliteration in identifying and converting the distinctive spellings between old jawi and modern jawi that had been modernized according to its roman form, while the former was more influenced by Arabic spelling [1]. For example it can be found in several studies like: [5] that suggested a rule based approach that involved stemming and filtering of old jawi source text, [6] which addressed one of the issues of letters error by converting a common character errors from ف into ڤ, and [7] which proposed vowel insertion rule for vocabs limitation before proposing a later study on another rule based approach for homograph [8]. Overall the limitation of these studies is similar in terms of requiring more additional rules, significant to the drawbacks of rule based approach of language dependent and resource consuming [4][9]. Despite of rule based importance to jawi studies, focusing too much on rule based approach could be one of the contributing factors to the slow-paced growth of jawi studies in NLP as statistical based is known to be more popular in machine translation studies up to these days.

For instance, one of the most powerful translation platform is Google translate, which is known to be utilizing this approach since 2007 [10]. Besides, the advantages of SMT include better accuracy as well as easier implementation compared to RBMT. Most importantly, it is trainable, adaptable and scalable as compared to RBMT

978-1-7281-3041-5/19/$31.00 ©2019 IEEE.

which is more rigid [4]. SMT in fact has been studied in both majority and minority languages widely. Despite of the focus on rule based, several studies on old jawi also have been emphasizing the need of statistical approach for old jawi. For instance [11] did an initial study of developing a jawi corpus with POS tagging by performing the tagging process on jawi [12]. [10] also suggested a hybrid model that combines rule based and statistical approaches such as Noisy Channel Model (NCM) or also known as Phrased Based SMT (PBSMT). Besides, in recent study by [12], the study focused on developing a Buckwalter jawi corpora that aims to work more efficiently in the tokenizing process of preprocessing level. Nonetheless, these studies are limited in identifying and emphasizing the significant effects of utilizing SMT for old jawi transliteration, in terms of its performance as well as the issues it may encounter.

Hence, the main aim of the research is to discover the significant effects of using a statistical approach on old jawi to modern jawi transliteration. In achieving such aim, two sub questions were formulated: How does PBSMT perform on old jawi to modern jawi transliteration measured by quality score of WER, and how does PBSMT affect the issues of rule based transliteration on old jawi to modern jawi such as vocals limitation and homograph, reduplication, letters error and combination of multiple words.

Lastly, the remainder of the paper was ordered as such: issues of rule based on old jawi transliteration like vocals limitation, homograph, reduplication, letter's error and combination of multiple words, then methodology of the study which discusses the research design as well as pilot study, followed by result, discussion and conclusion respectively. The reference can be found at the last section of the paper.

II. Issues of Rule Based on Old Jawi Transliteration

A. Vocals Limitation and Homograph

As old jawi was more influenced by Arabic and English at times, vocals usage was limited hence causing the difficulties of reading old jawi text. Consequently, it also affects the conversion of old jawi directly to roman, which becomes more difficult as compared to modern jawi to roman as the identification of the actual meaning or spelling of the word is hard to be defined in rule based [1][2][7]. Most importantly such issue caused to another issue, homograph. According to *Collins English Dictionary* [7], words that have similar spelling and may be similar in sound yet bear different meaning is defined as homograph

B. Reduplication

There are three types of reduplication in Malay language such as full reduplication, partial reduplication (vowel alternation, onset alternation, rhyming words) and echoic expression. Full reduplication is the repetition of the same exact word which is usually separated by dash or represented by an addition of '2' at its end. For instance *berlari2, sama-sama* or *hari2*. Whereas partial reduplication is adding a partial presentation of the word then combine it with the root word, for instance from *daun* to *dedaun,* which *de* is added to its root word *daun*. Besides, partial reduplication is also represented by the repetition of the root word yet the root word's phonemes or vowel could be altered. For example is *campur-baur* or *paku-paka*. Reduplication can derive several different meanings, not simply denoting the word as plural. Lastly for echoic expression, reduplication is used to add expressive effect like *ketuk-ketuk* and *anak-pinak* [14]. Significantly for rule based transliteration on reduplication of old jawi, the rule must incorporates all these types of reduplication in order to produce accurate transliteration. As the spellings are uncertain, more rules are needed which can lead to an endless effort in developing a machine transliteration for old jawi text. Moreover, the disadvantages also include high engineering cost, highly dependent on language used, not to mention resource consuming in terms of language expertise. Reduplication issue can be found in the study of old jawi transliteration using e-jawi [1][15].

C. Letter's Error

Letter's error is a common issue of old jawi conversion to another form especially for jawi characters which were not derived from Arabic letters but was represented by it on MM's writing like ف was used instead of ڤ or ك instead of ک or و instead of ۏ. As a result, rule based transliteration tool encountered difficulties in producing accurate transliteration, as it cannot differentiate whether the character in the source term is correctly used or not. However, letter's error is easier to be addressed compared to vocals limitation and homograph. This issue can be found in encountered in research done by [1][15][12]

D. Combination of Multiple Words

This issue can be found in the study of [1][10][15]. In this case, two or more words were combined together in shortening the writing time [10]. For instance the word مڠيكوتدي is the merging of two words: مڠيكوت and دي. The transliteration of this word will take a lot of rules to be defined for rule based approach as the term does not exist in Malay dictionary.

III. Methodolgy

The research method used was mix method and the strategy used was archival. Quantitative method was used in computing the quality of SMT output whereas qualitative method was utilized in assessing the issues of the output. For the archival strategy, the researcher gathered the archival records of Malay manuscripts as its sample data. Whereas for the sample data, a MM titled *Bidāyat al-Mubtadī bi-Faḍlillāh al-Muhdī* was chosen with a total of 885 sentences. The MM discusses on faith, prayer and fasting [16].

A. Jawi-Roman Transliteration Tool: E-Jawi version 3

E-Jawi is a transliteration tool accessible on the web that can be used to transliterate both roman to jawi and vice versa. The transliteration techniques used are purely Rule Based approach, which are algorithm and dictionary based. According to the developer, the tool was developed based on research work TERUJA [17] which focused on transliteration of modern jawi. Nevertheless, for the pilot study of this research, E-Jawi was used as a benchmark in testing the performance of Rule Based approach for both old jawi and modern jawi transliteration. The relevance is to identify and emphasize on the issues of transliterating old jawi texts compared to modern jawi.

For the sample data of old jawi texts, similar MM was used which is *Bidāyat al-Mubtadī bi-Faḍlillāh al-Muhdī*.

The total tokens for the dataset was 406, excluding of Arabic words. The transliteration was performed manually by transliterating each sentence using the web tool. Afterwards, each sentence was separated into tokens and was analyzed. Based on the result, several issues were found similar to other studies done on old jawi transliteration such as vocals limitation and homograph, reduplication, letters error and combination of multiple words.

B. Research Design

For the transliteration, this research is only focused on transliteration of old jawi to modern jawi. In previous research on jawi-roman transliteration done by [7], this transliteration process was an important part of RBMT, before transliterating the result to roman as shown in figure 1. However this research focused on old jawi to modern jawi as the transliteration of modern jawi to roman is easier compared to transliteration of old jawi [1]. Besides, the study on modern jawi to roman transliteration has been covered more extensively in past works.

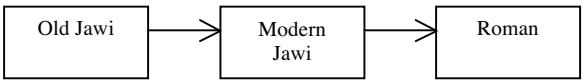

Figure 1 Transliteration process of old jawi to roman.

The SMT model chosen for this study was Phrase Based Statistical Machine Transliteration (PBSMT). There were three crucial aspects of performing PBSMT on old jawi to modern jawi, such as aligner, language model and decoder [2]. Aligner is the process of aligning the corpus into parallel corpus, in other words aligning the old jawi text to its modern jawi transliteration. Besides that, trigram was used for language model in calculating the probability of the target phrase. Lastly, stack decoder was used to compute probability based on best hypothesis. Stack decoder limits the best hypotheses through histogram and threshold pruning [18]. In this research, NLTK stack decoder was used. Overall there were two main steps involved: preparation of training dataset as shown in figure 2 and transliteration of training dataset as shown in figure 3.

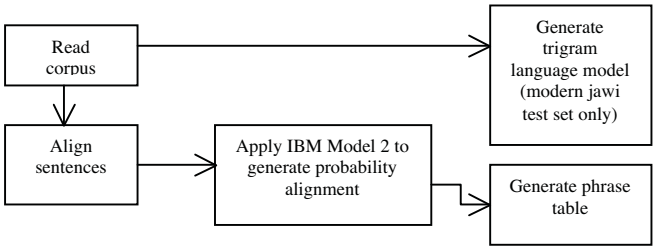

Figure 2 Preparation of training dataset for stack decoder.

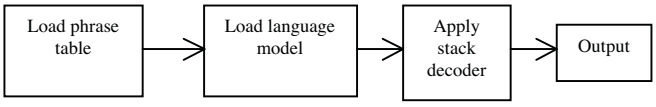

Figure 3 Transliteration of training dataset of old jawi to modern jawi using phrase based SMT.

For the analysis of the first research question, word error rate (WER) score was used in analyzing the quality score of the transliteration output. WER formula was computed using Python's library called *jiwer*. *Jiwer* is a simple package for calculation of word error rate between ground truth sentences and hypothesis. The formula of WER is depicted as following:

$$\textit{Error Rate} = (W/N) * 100$$

W implies the number of falsely transliterated words produced by the model compared to the validated human transliteration, whereas N is the total number of words being tested [19].

Whereas for the second research question, the outputs were assessed and the applicable outputs were classified according to vocals limitation and homograph, reduplication, letters error and combination of multiple words. The significance is to study the effect of PBSMT towards rule based issues as mentioned in Introduction section.

IV. RESULT

TABLE I. OVERALL STATISTICS OF TRAINING AND TEST DATASET

Title	Sentences	Token:Vocab		Training Set	Test Set
		O	M		
Bidāyat al-Mubtadī bi-Faḍlillāh al-Muhdī	885	8,568 :1,530	8,831 :1,432	700	185

* O = Original / Old Jawi * M = Modern Jawi

A total of 700 sentences were used in training the model, equal to 79% of the overall dataset and the remaining 185 sentences were used as test dataset. Besides that for the whole dataset, the total number tokens of old Jawi is 8,568, whereas for its modern form is 8,831. For vocabulary, old text corpus has 1,530 unique words whereas modern text corpus has 1,432 unique words. The statistics can be referred in table 1. Whereas for the length of the sentences, old jawi has longer sentence length as compared to modern jawi, which is 16 to 15 as depicted in figure 4. Besides, the figure also shows that most of the old jawi texts and its transliterations have similar lengths.

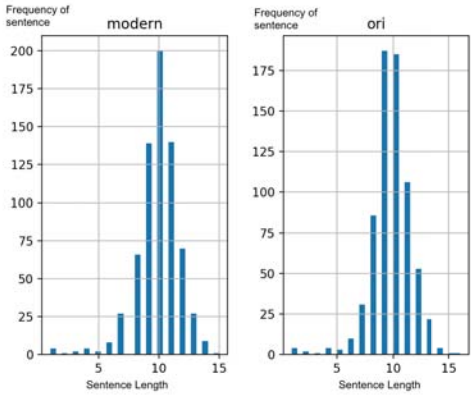

Figure 4 Length of sentences in the training dataset.

For the first research question, the research aims to assess the performance of PBSMT's output on old jawi to modern transliteration. Based on the result, all of the sentences transliterated had missing words. Example of the transliterated sentences was shown in figure 2, which shows that the transliteration results were partial. Besides, in figure 3, the transliterated text in modern jawi shows various lengths of sentences compared to original text, which means

that the lengths were mostly shorter than its original form. Significantly referring to table 2, the total number of differences between original text's tokens to transliterated tokens was 1,000 and the total number of differences for vocabs was 395. These results apparently impacted the WER of the output, which score is 0.66007483.

Original Text (Old Jawi)	Transliteration (Modern Jawi)
دالم إيت أمت قد مڠمبيل أير سمبهيڠ	دالم حدث جو دافون سنة قد مڠمبيل أير سمبهيڠ ايت امة
بسم الله دان ممباسوه تاڠن هيڤ	بسم الله في اوله واخره دان مماسه كدو تافق تاڠن هڤك
قد أڠكوتان أتاو تياد مك واجب قد سڬالا أڠكوتان	فد اڠكوتان اتو تياد مك واجب اي مرتاكندي فد سكل اڠكوتان
دان مهلاڠهلاڠ ممبايوه سڬالا آنق جاري	جار كانن در باوه داكو دان مهلاڠ2 سكل انق جار تاڠن كدو دڠن
باسوه يڠ قد كدو تاڠن هيڤ سمڤاي دان قد كدوا كاكي	باسه يڠ قد كدو تاڠن هڤك سمفي قد كدو لڤن دان قد كدو كاك

Figure 5 Sample of old jawi to modern jawi transliteration using PBSMT.

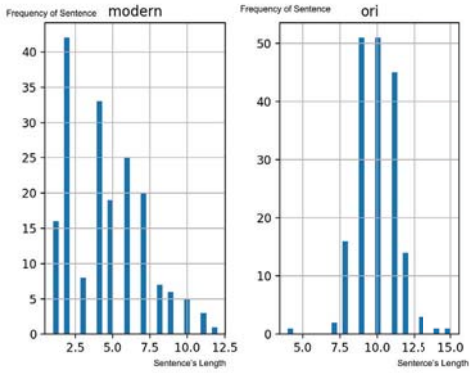

Figure 6 Length of sentences of the original text and its transliteration result in modern jawi.

TABLE II. OVERALL RESULTS OF TEST SET

Description	Result
Original text tokens (old jawi)	1,846
Transliterated tokens (modern jawi)	846
Total differences of tokens	1,000
Original text's vocabs (old jawi)	595
Transliterated vocabs (modern jawi)	200
Total differences of vocabs	395
WER	**0.66007483**

Other than that, there were several issues of old jawi transliteration that the research aims to observe on its occurrence in the output of PBSMT such as vocals limitation and homograph, reduplication, letters error and combination of multiple words. The results are shown in figure 7, 8, 9, 10 and 11 respectively.

Original Text (Old Jawi)	Transliteration (Modern Jawi)
كهيدڠ	كهيدوڠ
مڠحضركن	مڠحاضيركن

ملينكن	ملاينكن
جاد	جادي
كلاڠة	كلاڠيت
كمه	كوموه
سكل	سڬالا
الهم	أوليهمو

Figure 7 Successful transliterations of words with limited vocals.

Original Text (Old Jawi)	Transliteration (Modern Jawi)	Actual Meaning
برسوك2	برسوكا2	برسوكي2
تاه	تاهو	تاهي

Figure 8 Homograph issues found for old jawi to modern jawi transliteration using PBSMT due to vocals limitation.

In figure 7, words with limited vocals and its correct transliteration were listed. Overall a total of 9 words were extracted as sample result. The sample was chosen based on the ambiguity of reading the words due to missing vowels. Nonetheless, the research also found a total of 3 homograph words with wrong transliteration as depicted in figure 8 due to vocals limitation.

Original Text (Old Jawi)	Transliteration (Modern Jawi)
تيڤ2	تياڤ2
بركومر2	بركومركومور
مهلاڠ2	مهلاڠهلاڠ
ملبيه2	ملبيهلبيه
لكلاك	لاكي2

Figure 9 Successful transliterations of reduplicated words.

Whereas for reduplication, a total of 4 words were extracted as depicted in figure 9. Both words with number and without number were transliterated correctly.

Original Text (Old Jawi)	Transliteration (Modern Jawi)
سكاليڤون	سكاليڤون
ايت امة	إيت أمت
ملينكن كارن	ملاينكن كران
هڤك سمفي كڤد	هيڤك سمڤاي كڤد
تيك	تيڬ
فيهق	ڤيهق
فرمڤون	ڤرمڤوان
مانشي	مانسي

Figure 10 Successful transliterations of words with letters error.

For letter's error, 8 samples of successful transliterated words by PBSMT were extracted as listed in figure 10. Significantly, the extracted samples exhibited the issues of common characters error like ف instead of ڤ, ش instead of س, ک instead of ڬ, as well as ة instead of ت.

Original Text (Old Jawi)	Transliteration (Modern Jawi)
بارڠسباكيڬ	بارڠ سباڬايڬ
دمكينلاك	دميكين لاڬي

Figure 11 Successful transliterations of words with two or more combinations.

Lastly for the last issue of multiple words combination, the successful transliterations were shown in figure 11 with a total of 2 words as example.

V. DISCUSSION

Based on the result, the utilization of PBSMT in this research had shown significant improvement on the issues of rule based transliteration, despite of the high score of WER due to the notable number of partially inaccurate transliterations. Significantly such score was greatly impacted by the limitation of corpus size. Besides that, PBSMT does not produce accurate transliteration according to word order like found in a study by [20]. Despite that, all the issues observed like vocals limitation and homograph, reduplication, letters error and combination of multiple words had shown significant output, except for homograph.

In fact, such error also was caused by the limited corpus size used in the study.

VI. Conclusion

Based on the result, SMT that utilizes PBSMT as its model seems to be a promising approach for old jawi to modern jawi transliteration. As statistical approach derives its output based on its language model and phrase table, the model learned and produced better results and less effort in transliterating old jawi. Significantly as found by the research, several issues of rule based transliteration can be solved using PBSMT and vice versa. However, homograph issue of differentiating similar original words into different transliterations still persists with SMT's implementation. Overall, the main factor to these issues was the size of the corpus. Hence more transliterated manuscripts and other old jawi texts shall be collected in developing a better corpus especially in training the language model as well as in generating the phrase table, so that old jawi transliteration can be improved. Afterwards, the corpus also can be enhanced into annotated corpora in enhancing the output quality. Significantly according to [11], POS tagging can be used in overcoming homograph issue.

Nevertheless according to [7], a hybrid model that combines SMT and RBMT is another potential solution to homograph issue. Based on the previous works, the analysis done were limited in proposing a hybrid model for old jawi transliteration, either for old jawi to modern jawi or to roman. Based on the issues encountered by this research in using SMT, RBMT could help SMT in solving OOV issue such as in correcting the terms into proper spelling of modern jawi before directly transliterating the word using rule based. According to different studies, a hybrid approach has several advantages like better accuracy [4][7][8]. Furthermore, this study solely focused on old jawi to modern jawi transliteration following the previous works that focused on rule based. Another approach is to directly transliterate old jawi to roman, as roman Malay has been covered more extensively in NLP area. Hence before developing such model, a research on direct transliteration from old jawi to roman using SMT shall be done.

References

[1] C. W. S. B. C. W. Ahmad, K. Omar, M. F. Nasrudin, M. Z. Murah, & S. M. Azmi, "*Isu-Isu Dalam Transliterasi Mesin Manuskrip Melayu Ejaan Jawi Lama Kepada Jawi Baru*," Seminar Penyelidikan Jawi dan Manuskrip Melayu, 2012.

[2] C. I. Davis, "Tajik-Farsi Persian Transliteration Using Statistical Machine Translation," In *LREC,* 2012. pp. 3988-3995.

[3] M. M. Kashani, E. Joanis, R. Kuhn, G. Foster, & F. Popowich, "Integration of an Arabic transliteration module into a statistical machine translation system," In *Proceedings of the Second Workshop on Statistical Machine Translation*, Association for Computational Linguistics, 2007. pp. 17-24.

[4] K. Kaur, & P. Singh, "Review of machine transliteration techniques," *International Journal of Computer Applications*, *107*(20), 2014.

[5] C. W. S. B. C. W. Ahmad, "*Penterjemah Jawi lama kepada Jawi baru,"* Sarjana Tesis Sarjana, Fakulti Teknologi Dan Sains Maklumat, Universiti Kebangsaaan Malaysia, 2007.

[6] A. B. Juhaida, "*Transliterasi Jawi lama kepada Jawi baru berasaskan grafem (kajian kes pada Hikayat Merong Mahawangsa)*," Sarjana Tesis Sarjana, Fakulti Teknologi Dan Sains Maklumat, Universiti Kebangsaaan Malaysia, 2008.

[7] C. W. S. B. C. W. Ahmad, K. Omar, M. F. Nasrudin, M. Z. Murah, & S. M. Azmi, "*Machine Transliteration Design for Old Malay Manuscript*." In 2nd International Conference on Machine Learning and Computer Science (IMLCS' 2013), Kuala Lumpur, 2013. pp. 23–26.

[8] C. W. S. B. C. W. Ahmad, "*Transliterasi Mesin Bahasa Melayu: Jawi Lama - Rumi Berasaskan Petua,*" International Conference on Artificial Intelligence in Computer Science and ICT (AICS 2013), 25 - 26 November 2013, Langkawi, Malaysia [Online]. Available: https://worldconferences.net/proceedings/aics2013/toc/papers_aics2013/A091%20-%20TRANSLITERASI%20MESIN%20BAHASA%20MELAYU.pdf. [Accessed: 27 March 2018].

[9] M. L. Dhore, R. M. Dhore, & P. H. Rathod, "Survey on Machine Transliteration and Machine Learning Models," *International Journal on Natural Language Computing (IJNLC)*, *4*(2), 2015. pp. 9-30.

[10] C. W. S. B. C. W. Ahmad, K. Omar, S. M. Azmi & M. Z. Murah, "Kerangka Awal Transliterasi Mesin Teks Jawi Pada Batu Bersurat Piagam Terengganu (BBPT)", *Seminar Batu Bersurat Piagam Terengganu (BBPT) 2012, 25-26 Mac 2012, Kuala Terengganu, Malaysia,* [Online]. Available: https://www.researchgate.net/publication/237100094_KERANGKA_AWAL_TRANSLITERASI_MESIN_TEKS_JAWI_PADA_BATU_BERSURAT_PIAGAM_TERENGGANU_BBPT/references [Accessed: 30 June 2019].

[11] J. A. Bakar, K. Omar, M. F. Nasrudin, & M. Z. Murah, "Part-of-speech for old Malay manuscript corpus: A Review," In *International Multi-Conference on Artificial Intelligence Technology,* Springer, Berlin, Heidelberg, 2013. pp. 53-66.

[12] J. A. Bakar, K. Omar, M. F. Nasrudin, & M. Z. Murah, "NUWT: Jawi-specific buckwalter corpus for Malay word tokenization," *Journal of Information and Communication Technology*, Vol. 15 no. 1, 2016. pp. 107-131.

[13] Yonhendri, "Transliterasi Rumi ke Jawi berasaskan petua," Sarjana Tesis Sarjana, Fakulti Teknologi Dan Sains Maklumat, Universiti Kebangsaaan Malaysia, 2009.

[14] S. Nadarajan, "A Crosslinguistic study of Reduplication," *The Arizona Working Papers in Second Language Acquisition and Teaching*, *13*, 2006. pp. 39-53.

[15] S. M. A. Razak, M. S. A. Seman, W. A. W. Y. W. Mamat, & N. H. N. M. Noor, "Transliteration Engine for Union Catalogue of Malay Manuscripts in Malaysia: E-Jawi Version 3," In *2018 International Conference on Information and Communication Technology for the Muslim World (ICT4M)*, 2018. pp. 58-63. IEEE.

[16] [iium.on.worldcat.org]. (n.d.). Retrieved from https://iium.on.worldcat.org/search?queryString=no : 850746968#/oclc/850746968

[17] Heryanto, Omar & Nasrudin, "Transliteration Engine Rumi to Jawi (TERUJA)", *Ddms.usim.edu.my*, 2018. [Online]. Available: http://ddms.usim.edu.my/handle/123456789/6222. [Accessed: 27-Mar- 2018].

[18] Nltk.translate.stack_decoder — nltk 3.4.3 documentation. (n.d.). Retrieved from https://www.nltk.org/_modules/nltk/translate/stack _decoder.html

[19] A. S. Kumar, P. Premchand, & A. Govardhan, "Morphological Cross Reference method for English to Telugu Transliteration," *International Journal of Artificial Intelligence & Applications*, *2*(4), 2011. pp. 13.

[20] W. P. Pa, Y. K. Thu, A. Finch, & E. Sumita, "A study of statistical machine translation methods for under resourced languages," *Procedia Computer Science*, *81*, 2016. pp. 250-257.