Machine Transliteration Design for Old Malay Manuscript

Che Wan Shamsul Bahri Che Wan Ahmad, Khairuddin Omar, Mohammad Faidzul Nasrudin, Mohd Zamri Murah, Sanusi Mohd Azmi

Abstract—Jawi script is a script that has the Arabic influence. In the past, these writings are widely used by the Malay community as well as foreigners who have diplomatic relations, business, missionary and such. At that time, the Malay language is the lingua franca of this region. So there are many Malay heritages such as manuscripts, religious books, letters, documents and other agreements in the Jawi script. There are significant needs to do the transliteration of the Jawi text on the materials to Malay Roman. Thus, research on machine transliteration will help the effort. Many researches in machine transliteration in the world for high-level language have been done; such as English, European, and Asia languages such as Chinese, Japanese, Korean and Arabic. However the research in the context of the Malay language is still lacking, especially those involving the Romanized transliteration of Jawi. Jawi writing is quite different from the Urdu and Arabic although they share the same characters. Modern Jawi uses more vowel than the old version. These papers discuss the previous studies related to machine transliteration of the Malay language and approaches that can be used to develop it.

Keywords—Jawi, machine transliteration, Malay, rule base

I. INTRODUCTION

Jawi writing today is actually a Malay writing with Arabic influences that have been used nearly 700 years ago. This is evidenced by the discovery of Terengganu Inscribed Stone, dated 1303 AD [1]. However, the old Jawi is different compare Jawi script today in terms of the use of the vowel, the writing techniques and the use of new letters. It is seen in line with the development of the Malay language itself began to include some foreign words and technical terms especially in the advancement of science and technology today.

New Jawi spelling system was introduced in (1986) with the name of Pedoman Ejaan Jawi Yang Disempurnakan (Guidelines of the Enhanced Jawi Spelling) is intended to meet the present the needs of the future. This system is the result of a formula written in the Jawi National Convention held in 1984 at Kuala Terengganu, Malaysia. Jawi spelling system is compiled by [2] as its basis. According to [3], the system involves five processes, which are: maintaining confirmed words, perfecting the imperfect, creating the non-existence, clarifying the vague, tidying up the loose. In this paper, we define old Jawi as Jawi spelling system before era of Za’ba (before 1949). According to [4] there is a 30% difference between the old and new Jawi spelling. The old and new Jawi spelling systems are differences as shown in Table I below.

II. PROBLEM STATEMENT

In the context of the Malay language, transliteration is used to change the Jawi spelling to Rumi or vice versa. The main problem in the transliteration is when there are no matching characters in target text [5]. Jawi spelling system and Rumi (Roman) spelling are two different spelling systems. Jawi spelling read and written from right to left, while the reverse spelling of Rumi [6].

In addition, old Jawi spelling were not using more letters in practicing the concept of economy in the use of Jawi character than modern today [7]. According to [8] Jawi spelling as long as it is not consistent (varies according to the author) and many occur ambiguous word compared to new Jawi spelling system. The problems mentioned above are main challenges in the development machine transliteration for old Jawi to Rumi.

III. RELATED WORK

Machine transliteration is an important matter in the application of natural language processing (NLP), especially in translating an entity name from one language to another [9].
Among the world’s fastest-growing language in its transliteration mechanism is English, European, Asian languages such as Chinese, Japanese, Korean, Arabic, Urdu, Hindi, Punjabi, Taiwan, Korea, Japan, China, Thailand and others. Among them are [10], [11], [12], [13], [14], [15] and [5].

In Malaysia, research on machine transliteration is long, but its development is quite slow particularly the Malay language as compared with foreign languages.

There are many studies in Malay machine transliteration have done such as by [16], [17], [18], [19], [20], [21] and [22]. All researches on Malay transliteration are related to modern Jawi, except study done by [19] and [20] are related to old Jawi.

Several researchers such as [16], [17] and [18], using a character mapping technique that matches directly ١ to a, ٣ to b, and so on. Therefore, there are many words that cannot be converted correctly in the machine transliteration process. This is because there are more than matches for some of the letters in the target text.

In contrast to previous studies, [21] use rule-based transliteration technique in the study of Rumi to Jawi transliteration. Each Rumi word is processed following the Jawi spelling patterns. First, the word is divided into syllables, and then the syllables are matched with rule for Jawi conversion into Jawi spelling. Based on [21] research, there are words that cannot be converted into Jawi precisely because some of the problems; loan words from Arabic and English, the difficulty in distinguishing e-taling and e-pepet vowel, and the difficulty to determine whether the word has glottis ف or ك.

In the Jawi vocabulary, several words classified as past law or exempt from the law. Several words are not subject to the rule of Jawi. These are among the difficulties faced by [21] because Jawi spelling system is so unique.

[19] and [20] study is different as studying old Malay manuscripts. [19] using a Kitab Nazam (old Malay Manuscript) that has been digitized in advance for the study. [19] uses stemming and filtering model which is to make the process of rooting and separation between old and new Jawi words based on a Jawi corpus of built. While [20] use a transliteration based on grapheme to the epic of the old Malay manuscript, Merong Mahawangsa. With grapheme method, [20] successfully converted letters ف to ف for old Jawi words, instead of using grapheme ٣ in Rumi, not ف. This is because most of the old Jawi script does not distinguish phonemes ف and ف. [20] can reduce the gap between the old and new Jawi.

Many other studies are more focused on Rumi to new Jawi transliteration. Transliteration for old Jawi to Rumi is different because it contains its own rules and methods. Many studies have been done on transliteration for new Jawi but not for old Jawi.

IV. CHALLENGES

In doing transliteration of the old Malay manuscripts, there are some challenges that are identified as they use old Jawi spelling. Several challenges are discussed here.

A. Homograph and Less Vowel

Homograph is referring to two or more words that have the same spelling but different meanings[23]. Homograph can also coincided with a homonymous (same sound), for example, the word mereka (to create) with mereka (they) which is similar in terms of raw sound[24]. In the Jawi script, most homograph word occurred due to Jawi writing system itself that only use four vowels (١, ١, ١, ١) compared with six vowels (١, ١) in Rumi writing.

There are more homograph problems in old Jawi spelling because the economic principle of the lesser use of the vowel [7]. Here is an example of the old Jawi sentences occurs homograph.

ايه فلكي توفي، تمييز توفي برسم انيح جوه د جوه
Ayah pakai topi, tembak tupai bersama En. Johar di Johor.
(Dad wear a hat, shoot squirrels with Mr Johar in Johor)

ايه مالك وطير، مفتكل اتبر درامين فن
Ayah makan gulai, sedangkan anak bermain guli.
(Dad eat curry, while children play marbles)

B. Spelling Not Consistency

In old Malay manuscripts, spelling varies used even in a same book for some words. It also varies spelling between authors for some words. There are classification of old Jawi writer, including educated people in religion and palace writers anonymous, scholars and the general public[8]. Some of these people are not educated about the method of Jawi writing. Thus they write not according to the rules, not presentable and uncertain sentences. For example, [8] identified is in the pamphlets of ilmu wafaq (wafaq knowledge), ilmu hikmat (wisdom knowledge), mantera (mantras), traditional medicine and others.

C. Sentence Structure

Most of the old Jawi spelling or old manuscripts do not use punctuation [25]. The beginning of a new sentence usually begins with the word bahawasanya (whereas), maka (therefore), lagi (again) as sentence delimiter signal, even though sometimes the words maka (then) and lagi (again) act as conjunctions. Word dan (and) also used at the beginning of a sentence. In addition, there is a hadith or Quranic verses that can be used as a delimiter sentence.

D. Spelling System

Most of the old spelling does not use a point on the letter ١ (ya) if in the end of the word. This is because many of the old books printed in Arabia [25]. In Arabic, the letter ١ (ya) does not use a point when at the end of a word. Unlike other Malay letters, for example ٢ (ca), ٢ (nga), ٢ (nya) that dotted the three and ٢ (ga) which is used in real-dotted one. Sometimes there is also the point of the letter is placed under ٢ (ga).

E. Arabic Loanwords

There are consumption foreign words that absorption Arabic word for a kitab (book) or manuscript was written in Mecca. The Arabic word is synonymous with Islam itself. Sometimes, the author of a kitab, for example Kitab Hidayah al-Salikin...
difficult to find a corresponding meaning in Malay vocabulary because the original book is written in Arabic, Kitab Bidayatul Hidayat written by al-Imam al-Ghazali. If there are words in the Malay language, the purpose to be served may not meet the original intent of the word. Therefore, most scholars who wrote the book more comfortable use Arabic loanwords as Table II because more accurate meaning and mix with the Malay culture.

V. ARCHITECTURE OF OLD MALAY TRANSLITERATION

Fig. 1 show proposed architecture for Old Malay Transliteration. Hybrid model are highlighted in the proposed architecture considering the unique old Jawi spelling itself. Each word in the old Malay manuscripts will search through the list of words before the character mapping conducted at no matching words or Out of Vocabulary (OOV).

For the words that are not on the list after the test, then that word will also be tested using a search engine. For example, the sentence “لا فتُّ وُلَد” is not available in the list, and then the on-line resource will be used by the API search engine connections. The highest results will be selected as an outcome of transliteration as shown in Table III.

Some rulings in old Jawi writing like Shift Law, Insert Law, and character vowel materialized method have to be taken into consideration in design of the old Jawi to Rumi transliteration. Methods involve character vowel materialized Kāf-Ga Law and out of Deranglu law (د ﺭ ﻭ ﺒ). This is important to consider because several laws are used in old Jawi spelling but is no longer used in new Jawi spelling. For example, the Shift and Insert Law no longer used in new Jawi spelling while Kāf-Ga Law and out of Deranglu Law (د ﺭ ﻭ ﺒ) still remain in use till today. Deranglu Law use full vowels for its each syllables.

VI. CONCLUSION

From the above discussion, we can see some of difference between old and new Jawi spelling system. Development work of machine transliteration for old Jawi to Rumi is not an easy job compared new Jawi to Rumi. The work becomes more complicated and complex because the characters used are different. The results obtained from the proposed design may be accurate depending on how far we can follow some law or rule discussed above.
ACKNOWLEDGMENT

Special thanks to the International Islamic University College Selangor (KUIS) for providing the funds to continue studies at PhD level under the Academic Staff Training Scheme KUIS (SLAK). Thanks also to Universiti Kebangsaan Malaysia (UKM) for fees sponsorship to attend this seminar.

REFERENCES