# SIMPLE RULES MALAY STEMMER

Syed Abdullah Fadzli, A Khairani Norsalehen, I. Ahmad Syarilla, Hassan Hasni, M Satar Siti Dhalila

Faculty of Informatics
Universiti Sultan ZainalAbidin (UniSZA)
21300 Kuala Terengganu

## ABSTRACT

Stemming is a morphological analysis that tries to associate variants of the same term with a common root form. It is important to improve recall and precision in IR systems. Malay word stemming is considered complicated compared with other languages because of its unique morphological structure. Many research in Malay stemming relies heavily on dictionary which needs higher processing cost and offers lower coverage. This paper presents a stemming approach called UniSZA stemmer which attempts to reduce dictionary dependencies and lower the processing cost by proposing 7 simple rules. Experimental results show that the approach produces higher compression ratio and processing speed compared to RAO and RFO methods.

## KEYWORDS

Information retrieval, Malay stemming.

## 1 INTRODUCTION

Retrieving relevant information from a corpus of stored information is imperative in information retrieval as the number of written documents keep increasing these days [1]. Human query usually has variants of a basic word. In order to enhance the effectiveness of retrieved information, conflation method comes into play. Conflation is a computational procedure which identifies word variants and reduces them to a single canonical form [2]. Conflation algorithm can be generally categorized into stemming algorithm (language dependent) and string-similarity algorithm (language independent).

This paper focuses on stemming algorithm. Stemming is a computational process of reducing a word from its derived form into its root term. Stemming is essential in information retrieval field which gives significant influence in the indexing and searching process of document collections.

A document is represented by a vector of words or also known as terms [3]. Terms with the same stem generally have similar meanings. Thus, terms are grouped together into a common form to increase precision and recall in retrieving relevant documents against a given query.

This paper is organized as follows: Section 2 discusses the related works. Proposed method will be presented in Section 3 followed by evaluation and analysis in Section 4. Section 5 sums up the paper with conclusions and future works.

## 2 RELATED WORKS

In stemming algorithm, words with the same root are reduced to a common form by stripping each word of its derivational and inflectional suffixes [4,3]. As an example, the word ‘connect’, can be derived into ‘connection’, ‘connective’, ‘connected’, and ‘connecting’. The variable part is the ending, or ‘suffix’. Taking these endings off is called `suffix stripping' or ‘stemming’, and the residual part is called the stem. Various research has been done on the development of English language stemming algorithms for text retrieval purposes [4,5,3,6]. Porter [3] algorithm has become one of the widespread convention in the information retrieval community [7]. In Porter [3], 60 suffixes are looked for in order to produce word variant conflation. The stemming process uses a successive removal of short suffixes and not a single removal of the longest possible suffix. This makes Porter algorithm much simpler compared to other algorithms.

For the purpose of information retrieval in English language and similar languages such as Slovene and French, it is sufficient to remove only the suffixes [3]. However, in the case of Malay words; the process is more complicated [8]. This is due to the fact that Malay affixes consist of four different types verbal elements as outlined below:

i. Prefix – attaches itself at the beginning of a word
ii. Suffix - attaches itself at the end of a word
iii. Infix - usually located in the middle of a word
iv. Prefix-suffix pair - more than one affix that are attached to a word at the same time and usually positioned before and after the root words (eg : a combination of two affixes or a combination of a prefix and a suffix)

On the other hand; not only all affixes have to be removed in proper order [9], understanding the variations in different aspects of the four affixes shown above are crucial in the attempt to produce an effective Malay stemmer.

Early research in Malay morphology analysis and stemming was the work done by Abdullah in 1992. Abdullah has developed an algorithm for morphology analysis process for the word in Malay language and used it in Malay text retrieval [10]. In 1993, the first Malay Language stemmer was developed by Asim Othman [11]. Asim's [11] stemmer used rule-based approach with 121 morphological rules that were arranged and applied in alphabetical order. The word to be stemmed is checked against these rules and after appropriate affixes have been removed; the stemmed word is then checked against a general dictionary published by Dewan Bahasa and Pustaka. However, when Asim’s [11] approach was applied to a selection of ten chapters of the Quran and ten researches abstract as test data, it produces many errors.

Among the weaknesses to Asim’s method were the limitations on the number of morphological rules, arrangement or order of implementation of the rules, and the use of general dictionary to check for root words. Analyzing these weaknesses, Fatimah [9] proposed an improved algorithm with that integrates rule-based approach (561 rules) and lookup against a root words dictionary to check the

stemmed word [9]. This approach is called Rules Application Order (RAO) since the rules references were adapted according to specific orders. Her algorithm has been tested against a set of words from chapters in the Holy Book of Qur'an and it was found that the RAO approach reached 61.35% compression rate [9].

In 1999, Zainab [12] has developed a combination of conflation methods using N-gram string similarity and RAO stemming algorithm to evaluate the retrieval effectiveness on Malay documents [12]. Further work on RAO stemming algorithm and conflation methods using N-gram was pursued by Sembok and Abu Bakar [13] in the retrieval of Malay documents where they reported improvement in retrieval effectiveness using combined search of N-gram matching and stemming.

Other versions of Malay stemmers have been reported by Sock, Cheng and Abdullah [14] and Idris [15] but only the work done by Ahmad [9] and Abu Bakar [12] reported about the retrieval performance on Malay information retrieval systems. On the other hand, Abdullah et. al had proposed an approach called Rules Frequency Order (RFO) that was developed based on RAO approach They claimed that their approach provides a higher percentage of stemming correctness compared to RAO [16]. Based on their evaluation, they reported that RAO produces 4.4% errors while using the RFO approach, the errors were only 1.4%.

Although the work on stemming Malay words had already started since the past 19 years, it seems that the area of research is largely unexplored. Starting from Asim's [11] algorithm, other pivotal work includes the RAO approach proposed by Fatimah [9] which then became the instigator to related works in Malay stemming.

The main problem with RAO stemmer is that, it relies heavily on a dictionary consisting of 22,481 Malay root words. Every time a rules is applied, the resulted word is checked against the dictionary to determine whether to stop or continue with the next rule, requiring a total of eight (8) times dictionary checking process. In addition, it also relies on a list of 432 variation of prefix, suffix and prefix-suffix pairs. Each word is compared against this list in order to identify its affixes. This approach requires high processing cost which is crucial when indexing huge amount of documents.

## 3 PROPOSED METHOD

The paper introduces a new Malay stemming method, called UniSZA stemmer, which tries to reduce this dependencies by offering 7 simple rules (including *check dictionary*). These rules are developed to identify and remove prefixes and suffixes in Malay words namely; *Check Dictionary, Check Length, Double Words, Prefix, Suffix, Change Spelling* and *Suffix-i*.

- *Check Dictionary* – Malay words are considered complicated when dealing with suffixes and prefixes. There are many Malay words which start with spelling identical to prefix and end with spelling identical to suffix, but they are neither prefix nor suffix. For instance, the word '*perhati*' which means to observe, starts with '*per*' which is a

typical prefix in Malay words. In this situation, it is very difficult to determine whether the spelling at the beginning of the word is a form of prefix or it is part of the root word. The easiest way is to compare the word with dictionary of Malay root words.

- *Check Length* - This function sets the threshold of the word length. According to empirical evidence, the threshold is set to minimum of 3 letters which means only words with 4 letters and above will be stemmed.
- *Double Words* – Sometimes Malay language use duplication of words to express plural. For example, the word '*bangunan*' is used to represent a building, while the same word is repeated twice to represent many buildings; '*bangunan-bangunan*'. In some circumstances, spelling of duplicated words are changed, for example '*gunung*' which represents mountain is duplicated into '*gunung-ganang*' to represents mountains (plural). This rule is developed to identify the root words of Malay words duplications.
- *Prefix* – This rule identifies word prefixes based on an enhanced Malay prefixes library which was developed based on RAO stemmer. The Malay prefixes library consists of 107 variants of prefixes for Malay words. Each word is compared against the library to identify any prefixes that should be removed from the spelling.
- *Suffix* – Using the same approach as *Prefix*, this rule will identify word suffixes based on a Malay suffixes library consisting of 97 variants. Any spellings matched against the library will be removed.
- *Change Spelling* – This rule tries to change the spelling of Malay words after it has been processed by *Prefix*. There are many occasions in Malay words where the spellings are changed after the addition of prefixes, hence the spellings must be restored after the removal of the prefixes. For example, the root word *tulis* means to write, when combined with prefix *pen*, will be spelled as *penulis* which means a writer. The first letter '*t*' in the original root word spelling is removed after the prefix being added.
- *Suffix-i* – Malay words have a special rule when dealing with suffix '*i*'. This rule states that only words that ends with two (2) consecutive vowels including the letter '*i*' at the end will be considered having a suffix of '*i*'. Therefore the suffix '*i*' will be removed. For example, the letter '*i*' in the word '*beli*', which means to buy, is part of the root word. While the word '*punyai*', which means having, contains the prefix '*i*' at the end of the spelling.

Based on the experiments done by Abdullah et al. [16], the arrangement of the rules gives significant impact to the performance of the stemming process. Following their approach, an experiment was conducted using 10 different documents, taken from 10 random *surah* of Quranic translation. Table 1 lists the details of each document. Each *surah* consist of different words count ranging from 2,312 words (*surah Maryam*) to 8,620 words (*surah Ali-Imran*). Total number of words in all 10 documents is 41,938.

Using the constructed rules, five (5) distinct rules arrangement were used and

compared in an experiments using the 10 documents. The five (5) distinct rules arrangements are:

- **Arrangement A:** *Double Words → Check Dictionary → Check Length → Suffix → Prefix → Change Spelling →*
- *Suffix-i*
- **Arrangement B:** *Double Words → Check Dictionary → Check Length → Prefix → Change Spelling → Suffix → Suffix-i*
- **Arrangement C:** *Double Words → Check Dictionary → Check Length → Prefix → Suffix → Change Spelling → Suffix-i*
- **Arrangement D:** *Double Words → Check Dictionary → Check Length → Suffix-i → Suffix → Prefix → Change Spelling*
- **Arrangement E:** *Double Words → Check Dictionary → Check Length → Suffix → Suffix-i → Prefix → Change Spelling*

**Table 1:** Experimental documents

| Document | Surah Name | No. of Words |
|---|---|---|
| 1 | Yunus | 4654 |
| 2 | Ali-Imran | 8620 |
| 3 | Al-Maaidah | 6637 |
| 4 | Al-Anfaal | 2959 |
| 5 | Al-Isra' | 3716 |
| 6 | Maryam | 2312 |
| 7 | Al-Hajj | 2930 |
| 8 | Az-Zukhruf | 2444 |
| 9 | Yusuf | 4241 |
| 10 | Al-Qasas | 3425 |
| *Total no. of words in all documents: 41938* | | |

**Table 2:** Comparisons of five (5) different rules arrangements

| Arrangement | No. of correctly stemmed words | Accuracy |
|---|---|---|
| **A** | 37649 | 89.77% |
| **B** | 37835 | 90.22% |
| **C** | 37847 | 90.25% |
| **D** | 37606 | 89.67% |
| **E** | 37603 | 89.66% |
| *Total no. of words in all documents: 41938* | | |

Stemming accuracy can be defined as the percentage of the number of correct stems against the size of document samples. Table 2 lists the comparison results between the five (5) arrangements. The results showed that the arrangement C is the best arrangement with highest value of accuracy of 90.25%.

Based on the experiment, arrangement C is then chosen as the best arrangement for the proposed method. The proposed arrangement can be described as follows:

***Step 1:*** Get the next word until last word;

***Step 2:*** Check if the word is a double words; if yes, choose the first word as the root word.

***Step 3:*** Check the word against dictionary; if it exist, the word is the root word and go to Step 1.

***Step 4:*** Check the length of the word; if it is less than 4 letters, the word is the root word and go to Step 1.

***Step 5:*** Check the word spelling against Prefix list; if matched, remove the prefix

***Step 6:*** Check the word spelling against Suffix list; if matched, remove the suffix

***Step 7:*** If prefix is removed in Step 5, check the beginning spelling; if missing letter is found, restore it

***Step 8:*** Check if the word have a suffix of *i*; if yes, remove it.

## 4 EVALUATION

In this paper, the evaluation of stemming algorithm is concerned with the compression and processing speed of the algorithm. Following Abdullah's evaluation approach [16], compression is measured by calculating the number of distinct roots produce by the algorithm. While the number of words correctly stemmed can measure the accuracy performance of the algorithm. At the end of each stemming process, the produced stem words is checked against the Malay root word dictionary to determine the existence. If the new stemmed word exist, it is considered as correct.

The Malay root word dictionary root word used in RAO algorithm is used in this evaluation process. RAO stemmer is run on the test data for the purpose of comparison. The test data consist of 10 documents providing a total of 41,938 words, with 3,512 distinct words.

The employment of UniSZA's stemming algorithm reduces the number of distinct words from 3,512 to only 1,150 distinct root words, or 67.26% compression rate, compared to only 61.4% for RAO and 62.3% for RFO using their own set of data. The result for each stemmer is shown in Table 3.

**Table 3:** Evaluation on Compression Performance

| Stemmer | Total Distinct Words in Documents | Distinct Stemmed Words | Compression |
|---|---|---|---|
| **RAO** | 6900 | 2667 | 61.35% |
| **RFO** | 6900 | 2602 | 62.29% |
| **UniSZA** | 3512 | 1150 | 67.26% |

UniSZA stemmer performs better compared to other algorithms in term of compression ratio. It is also much higher than compression achieved by other language stemmer, English 26.2% to 50.5% and Slovene 54.7% [16].

The evaluation on speed performance is done by comparing the processing speed between UniSZA stemmer with RAO stemmer. Both algorithm is applied on the same document samples and the same environment for 10 times, and the average processing speed is calculated. The result is shown in Table 4.

**Table 4:** Average Processing Speed in Milliseconds

| Document Samples | Average Processing Speed (milliseconds) | |
|---|---|---|
| | **UniSZA** | **RAO** |
| **1** | 23.4 | 46.8 |
| **2** | 41.5 | 76.9 |
| **3** | 32.9 | 61 |
| **4** | 15.4 | 35.9 |
| **5** | 17 | 42.4 |
| **6** | 15.8 | 31.5 |
| **7** | 15.5 | 31.1 |
| **8** | 15.6 | 31.1 |
| **9** | 20.1 | 48.5 |
| **10** | 18.7 | 36 |
| **Average** | **21.59** | **44.12** |

The processing speed achieved by RAO stemmer is 51% slower than UniSZA stemmer due to the excessive use of

dictionary. UniSZA stemmer with much simpler rule set including 1 dictionary check process are able to increase the algorithm speed significantly.

## 5 CONCLUSION AND FUTURE WORK

The proposed Malay stemming algorithm focuses on improving the processing speed of available methods. Not only it produces higher compression ratio, the evaluation results also shows that the 7 simple rules proposed are able to produce significant processing speed compared to existing methods. However, this ongoing research still rely on Malay root dictionary to maintain the stemming accuracy. Future works will involve introducing additional rules which could eliminate the dictionary dependencies, hence improving the processing speed.

## REFERENCES


[1] T.M.T Sembok, "Character Strings to Natural Language Processing in Information Retrieval," in *Proceedings of the 6th Intrenational Conference on Asian Digital Libraries, ICADL*, Kuala Lumpur, 2003.

[2] F Ahmad, M Yusoff, and T.M.T Sembok, "Experiments with A Malay Stemming Algorithm," *Journal of American Society of Information Science*, 1996.

[3] M. F Porter, "An Algorithm for Suffix Stripping," *Program*, vol. 14, no. 3, pp. 130-137, 1980.

[4] J.B. Lovins, "Development of a Stemming Algorithm," *Mechanical Translation and Computational Linguistics*, vol. 11, no. 1-2, pp. 22-31, 1968.

[5] A. Margaret, Hafer and F. Stephen Weiss, "Word Segmentation by Letter Successor Varieties," *Information Storage and Retrieval*, vol. 10, no. 11-12, pp. 371-385, 1974.

[6] M Lennon, D Peirce, B Tarry, and P Willett, "An Evaluation of Some Conflation Algorithms for Information Retrieval," *Journal of Information Science*, vol. 3, pp. 177-188, 1981.

[7] Donna Harman, "How Effective is Suffixing?," *Journal of The American Society for Information Science*, vol. 42, no. 1, pp. 7-15, 1991.

[8] T.M.T Sembok, M Yusoff, and F Ahmad, "Characterization of An Experimental Collection Using Malay Language," in *4th International Conference and Exhibition on Multi-Lingual Computing ICEMCO-94*, Cambridge: University of Cambridge Centre of Middle Eastern Studies, 1994.

[9] F Ahmad, "A Malay Language Document Retrieval System: An Experimental Approach and Analysis," Universiti Kebangsaan Malaysia, Bangi, 1995.

[10] M.T Abdullah, "Sistem Bantuan Pembinaan Kamus Berasaskan Pangkalan Data Teks Bebas," Universiti Teknologi Malaysia, Kuala Lumpur, 1992.

[11] A Othman, "Pengakar Perkataan Melayu Untuk Capaian Dokumen ," Universiti Kebangsaan Malaysia, Bangi, 1993.

[12] Z Abu Bakar, "Evaluation of Retrieval Effectiveness of Conflation Methods on Malay Documents," Universiti Kebangsaan Malaysia, Bangi, 1999.

[13] T.M.T Sembok and Z Abu Bakar, "Characteristics and Retrieval Effectiveness of n-gram String Similarity Matching on Malay Documents," in *10th WSEAS International Conference on Applied Computer and Applied Computational Science (ACACOS '11)*, Venice, 2011.

[14] Y.T Sock, S.O Cheng, and N. A Abdullah, "On Designing an Automated Malaysia Stemmer for the Malay Language," in *5th International Workshop Information Retrieval with Asian Language*, 2000, pp. 207-208.

[15] N Idris, "Automated Essay Grading System using Nearest Neighbour Technique in Information Retrieval," Universiti of Malaya, Kuala Lumpur, 2001.

[16] M.T Abdullah, F Ahmad, R Mahmod, and T.M.T Sembok, "Rules Frequency Order Stemmer for Malay Language," *IJCSNS International Journal of Computer Science and Network Security*, vol. 9, no. 2, pp. 433-438, 2009.

[17] T.M.T Sembok, Zainab Abu Bakar, and Fatimah Ahmad, "Experiments in Malay Information Retrieval ," in *2011 International Conference on Electrical Engineering and Informatics*, Bandung, Indonesia, 2011.

[18] T.M.T Sembok, "Word Stemming Algorithms and Retrieval Effectiveness in Malay and Arabic Documents Retrieval System," *World Academy of Science, Engineering & Technology*, vol. 10, pp. 95-97, 2005.