Rules Frequency Order Stemmer for Malay Language

**Muhamad Taufik Abdullah†, Fatimah Ahmad†, Ramlan Mahmod†
and Tengku Mohd Tengku Sembok††**

Faculty of Computer Science and Information Technology, Universiti Putra Malaysia,
43400 UPM Serdang, Malaysia

Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia,
43600 Bangi, Malaysia

Summary
The importance of stemmer is obvious with the advent of effective information retrieval systems. Unfortunately, Malay stemming problems are difficult to solve due to complexity of words morphology. The Rules Application Order (RAO) stemmer is examined for enhancing performance to minimize the percentage of stemming errors. This paper presents a stemming approach called Rules Frequency Order (RFO). RFO rearranges the stemming rules according to the frequency of their usage from the previous execution. It shows that the approach provides a higher percentage of stemming correctness as compared to RAO stemming approach.

Key words:
stemming, stemmer, information retrieval, Malay language.

1. Introduction

A stemming algorithm is defined as a computational procedure that will reduce all the inflectional derivational variants of words to a common form called the stem [1]. For example, the *words group, groups, grouped, grouping*, and *subgroups* are conveyed to the stem *group*. Precisely, the stem of a word is obtained by removing all or some of the affixes attached to the word.

This algorithm is regarded to be very useful for varying reasons especially in the fields of information retrieval and computational linguistics. In information retrieval, grouping words having the same stem (or indexing term) will increase the success of matching of documents to a query [2,3], reduce the size of index files, and increase the speed of retrieving. Whereas, in computational linguistics, linguists are more interested in finding a linguistically correct stem. It seems that the affixes are more important to them than the stem itself since the affixes represent information about the grammatical function of a word, and thus helping the analysis of a sentence syntactically.

There is a need for an effective Malay stemming algorithm in building an effective information retrieval system. The existing Malay stemming algorithm developed by Ahmad [4] has produced many errors when applied to the test data, a set of the Malay translation of the first two chapters in the Quranic, which contains 2116 unique words.

In this paper, an approach called Rules Frequency Order that provides a higher percentage of stemming correctness will be proposed. This technique is developed based on Rules Application Order approach.

This paper is organized as follows: In section 2, the related works will be discussed. In section 3, the Malay affixes are presented. In section 4, the proposed Malay stemming approach will be discussed. In section 5, the performance of the proposed Malay stemmer is analyzed in terms of percentage of errors, and a comparison with RAO stemmer is given. We draw our conclusions in section 6.

2. Related Works

Research in the Malay language text retrieval systems has been left lagging behind as compared to research that has been done on other languages such as English and other European languages. Stemming algorithms available for Malay information retrieval systems are not many. Abdullah [5] has developed an algorithm for morphology analysis process for the word in Malay language and used in Malay text retrieval. Othman [6] has also developed a stemming algorithm for Malay language using rule-based approach and a dictionary to check the stemmed word. Later, Ahmad [4] has developed a Malay stemming algorithm and a stemming approach named the Rules Application Order approach. Abu Bakar [7] has developed a combination of conflation methods of N-gram string similarity and stemming. Another version of Malay

Manuscript received February 5, 2009
Manuscript revised February 20, 2009

stemmers are also reported by Sock et al. [8] and Idris [9]. However, only the experiment done by Ahmad [4] and Abu Bakar [7] have reported about the retrieval performance on Malay information retrieval systems.

3. Malay Affixes

The Malay affixes consist of four different types, which are the prefix, suffix, prefix-suffix pair, and infix. Unlike English stemmer which work quiet well just by removing suffixes alone to obtain the stems, an effective and powerful Malay stemmer not only must be able to remove the suffixes, but also the prefixes, prefix-suffix pairs, and infixes as well [4]. Without removing all these affixes, the stems cannot be effectively used to index Malay documents. The question arises here is that which is the best order of the affixes to be applied in term of producing the minimum number of stemming errors. We test all the possible combinations of the affixes to the Malay translation of the first two chapters in the Quranic.

With the four different types of affixes, therefore there will be 24 different combinations. We compare the three most frequently used affixes, and maintain the fourth affix as the last in the order of affixes in order to reduce the number of different combinations to 6. We count the frequency of all affixes on the test data. The test data consist of 2116 unique words, which contain: 209 stop words; 568 root words; 310 prefixes; 376 suffixes; 504 prefix-suffix pairs; 0 infix and 149 reduplication words. It can be seen that there is no presence of infixes at all for the test data. It can be concluded that the infix will always be the last in the list of six different combinations. The six combinations will be called Test 1, Test 2, Test 3, Test 4, Test 5, and Test 6. We added another test to compare with a list of all affixes sorted in alphabetical order; this test is called Test 7. The tests are as follows:

Test 1:	pr-ps-su-in
Test 2:	pr-su-ps-in
Test 3:	ps-pr-su-in
Test 4:	ps-su-pr-in
Test 5:	su-pr-ps-in
Test 6:	su-ps-pr-in
Test 7:	alphabetical

Where:

pr:	the prefix rules
su:	the suffix rules
ps:	the prefix-suffix rules
in:	the infix rules
alphabetical:	the alphabetical order of all rules

4. The Malay Stemming Approach

In order to develop the actual stemming algorithm, it is important to have a list of affixes for the language first. The list will help in the understanding of Malay words, their morphological rules and gives an insight reflection of the structure of the Malay words that will be stemmed. The list of affixes that is created by Ahmad [4] is been considered to be included in the preliminary list.

In order to determine the most effective stemming algorithm for the Malay language, a few experiments need to be performed. We start with a list of affixes called set A; consist of 418 rules, and the spelling variations rules developed by Ahmad [4].

All the seven combinations of the rules application order need to be tested to determine which group of rules to trigger first to produce minimum number of stemming errors. Since the Malay stemmer is going to be used on the Quranic collection, therefore the first two chapters of the Quran are taken for these particular experiments. The rules being used here are from set A, which contain a limited number of rules that exclude modern and scientific affixes. A dictionary of Malay root words is needed to ensure that the root words obtained are valid ones. We used the dictionary of root words from SISDOM98, which contain 22,429 entries. We also used the list of 314 stop words developed by Ahmad [4].

When applying a rule in stemming a word, a minimum length constraint on the root obtained is imposed. Ahmad [4] suggested for the prefix and suffix rules, the minimum length of the root word is 2 whereas for infix and prefix-suffix pair rules, the minimum length is 3. For example, the minimum length constraint of 3 for the prefix-suffix pair procedure is chosen in order to correctly stem the word such as *bersalah* to *salah*. If a minimum length constraint of 2 is chosen, it will be wrongly stemmed to *sa* which is appears in the root dictionary.

The number of words wrongly stemmed can measure the performance of the algorithm. We run Ahmad's algorithm on the test data for the purpose of comparison. This stemming process is called Rules Application Order (RAO). The results obtained are listed in Table 1. From the results shown, it can be concluded that the second test which performs the checking of the prefix first, followed by the suffix, then checking of the prefix-suffix pair and finally checking of the infix is the best method.

We added another eight affixes to the set A. In addition, some modifications have been made to the spelling variations rules. For the dictionary of root words, we added a few entries. With that, the new root words dictionary contains 22433 entries. Then, we run Ahmad's [4] algorithm again on the test data. This stemming process is called Rules Application Order 2 (RAO2). The results obtained are listed in Table 1. From these results, the number of words wrongly stemmed is reduced and Test 2 gives the least number of errors compared to all the other tests. Therefore, we can conclude that the list of rules, spelling variations rules, and root words dictionary give affect the performance of stemming algorithm.

Further, some modifications have been made to the Ahmad's [4] algorithm and will be explained in the next section. Then, we run the new algorithm on the test data. This stemming process is called New Rules Application Order (NRAO). The results obtained are listed in Table 1. The results shown the number of words wrongly stemmed are reduced and Test 2 gives the least number of errors compared to all other tests. These results indicate that the new stemming algorithm performs better than Ahmad's algorithm.

Furthermore, using NRAO algorithm and all rules from NRAO, we run another experiment. However, the rules are sorted in decreasing order according to the frequency of each rules applied in NRAO. This stemming process is called Rules Frequency Order (RFO). The results obtained are listed in Table 1. From these results, the number of words wrongly stemmed for all tests are reduced and the least number of errors occur in Test 1, Test 2, Test 5, and Test 7. These indicate that RFO stemmer is better than the previous stemmer.

Table 1: Comparison between the stemmer

	Number of words wrongly stemmed			
Test	**RAO**	**RAO2**	**NRAO**	**RFO**
1	105	78	54	30
2	93	63	38	30
3	106	80	56	32
4	107	83	59	32
5	94	66	41	30
6	94	67	42	31
7	105	78	52	30
Number of rules	418	426	426	426
Minimum errors	93	63	38	30
Maximum errors	107	83	59	32

As the summary of the experiments in this section, Table 1 shows the comparison among the seven experiments for the stemmer. As can be seen in Table 1, the minimum errors are occur in Test 2 which perform the checking of prefix, suffix, prefix-suffix pair, and infix. Whereas, experiment by Ahmad [4] found the minimum errors in test which perform the checking prefix, prefix-suffix pair, suffix, and infix. From this, we can conclude that the best order of the affixes to be applied in term of producing the minimum number of stemming errors is depend on the document collection. For our collection, the best order of the affixes is prefix, suffix, prefix-suffix pair, and infix.

Comparing the results obtained from RAO stemmer, the RAO2 stemmer seems to be superior. Whereas NRAO stemmer is superior compare to RAO2. While RFO stemmer performs better than NRAO. In term of the number of minimum errors produced by the seven stemming process, it can be seen that RFO produced the least number of words wrongly stemmed from all tests. Whereas, for the number of maximum errors, Table 1 also shows that RFO stemmer produced the least number of maximum errors from all tests. Therefore, we can conclude that RFO stemmer is the best stemmer in term of the least number of words wrongly stemmed.

The performance of the stemmer can be improved by adding a few appropriate affixes into the list of rules, modifications of the spelling variations rules, and adding a few missing words into the dictionary of root words. From the results, it has reduced the maximum number of stemming errors from 107 for RAO to 83 for RAO2. The modifications of stemming algorithm have reduced the maximum number of stemming errors from 83 in RAO2 to 59 in NRAO. Whereas the performance of the stemmer also can get further improve by rearrange the rules. The list of rules is sorted in decreasing order according to the frequency of rule's usage in previous stemming. As the results shown, the maximum number of stemming errors has reduced from 83 for NRAO to 32 for RFO.

Comparing the results obtained from Test 1, Test 2, Test 3, Test 4, Test 5, and Test 6, it can be seen that Test 7 performs among the worst case for RAO, RAO2, and NRAO stemming process. However, when we used RFO stemmer, Test 7 performs as good as the best cases, that are Test 1, Test 2, and Test 5.

The types of error produced by RFO stemmer for all the seven tests are given in Table 2. RFO stemmer produced the same error for the same word in this test. Altogether there are 33 unique errors. We classify errors into five types namely, overstemming, understemming, spelling exception, and others.

The distributions on the number of unique errors for all seven tests using RFO stemmer are listed in Table 3. The errors caused by type understemming constitute the most number of errors. This is due to the order in which rules are applied.

The stemming algorithm consists of two main parts: a basic stemming procedure and the recoding procedure where both are being employed within the algorithm as follows.

The basic stemming procedure consists of the initial checking of the dictionary where every input word is being checked with the dictionary of root words. If the word exists in the dictionary, then the word is immediately taken as a root word. Otherwise, the word has to undergo the stemming process. First, a pass is made through the list of affixes, and if a match is encountered, a deletion is made. After deletion, the stemmed word will be checked against the dictionary to make sure it is a valid root. If it exists in the dictionary, then it will be taken as a root word, otherwise the recoding procedure that handles spelling exceptions and variations will be performed. The dictionary is again checked and if it still does not exist, the next rule on the list will be triggered. All these processes will be repeated until the rule is applied. Finally, if the word cannot find a match in the list of morphological rules or the stemmed word does not exist in the dictionary, the original word will be returned. The stemming approach that has been taken in this stemming algorithm is based on RAO [4] approach.

In this stemming algorithm, the rule-based approach is employed in order to make the computer program flexible in accommodating changes in the morphological rules. A set of rules which defined prefixes, suffixes, prefix-suffix pairs, and infixes are written in the following formats:

Prefix rules format: prefix+
Example: *be+*
Suffix rules format: +suffix
Example: *+an*
Prefix-suffix pair rules format: prefix+suffix
Example: *ber+kan*
Infix rules format: +infix+
Example: *+er+*

To find the right root, a context-sensitive approach that adds constraints to the basic stemming procedure needs to be employed. Two general types of constraints are being used in this algorithm:

i. Quantitative constraints:
The minimum stem length, which should be left after the removal of the affix, must be determined. The minimum length constraint for the prefix and suffix rules is 2. For the prefix-suffix pair rules and infix, the minimum rules constraint is 3.

ii. Recoding rules:
The recoding procedure takes place where the spelling rules must be used to improve the accuracy of conflation of the roots produced by the stemming algorithm. Spelling exceptions and variations are not declared using the morphological rules but are handled in the program itself because of its complexity. The spelling exceptions and variations only apply on some of the prefixes, prefix-suffix pairs, and suffixes where some of the first letters of root words need to be dropped when combined with these affixes.

The spelling exceptions for prefixes involved in this stemming algorithm are listed in Table 4 and the spelling exception for suffix is shown in Table 5.

Table 4: Spelling exceptions for prefixes

Prefixes	The first letter of root word to be dropped
mem, pem	f, p
meng, peng	k
meny, peny	s
men, pen, sepen	t

Table 5: Spelling exceptions for suffix

Suffix	The last letter of root word to be dropped	Change the last letter to
an	p	b

The spelling variations rules list the Malay affixes with their corresponding first letters of root words, which are allowed to be attached to them. The notation used to describe this morphological rule is given as follows:

men+ c,d,j,sy,t,z

Which means that prefix *men* is used with words beginning with *c, d, j, sy, t,* and *z*. There is a list of 40 new spelling variations rules used in this stemming algorithm.

The new Malay stemming algorithm can be described as follows:

Step-1: Get the next word until the last word;
Step-2: Check the word against the dictionary; if it appears in the dictionary, the word is the root word and goto Step-1;
Step-3: Get the next rule; if no more rules available, the word is considered as a root word and goto Step-1;
Step-4: Apply the rule on the word to get a stem;
Step-5: Perform recoding for prefix spelling exceptions and check the dictionary;
Step-6: If the stem appears in the dictionary, the stem is the root of the word and goto Step-1; else goto Step-7;
Step-7: Check the stem from Step-4 for spelling variations and check the dictionary;
Step-8: If the stem appears in the dictionary, the stem is the root of the word and goto Step-1; else goto Step-9;

Step-9: Perform recoding for suffix spelling exceptions and check the dictionary;

Step-10: If the stem appears in the dictionary, the stem is the root of the word and goto Step-1; else goto Step-3;

5. Evaluation of the Stemmers

Evaluation of stemming algorithm concerns with the correctness of algorithm itself. This type of evaluation can carried out by seeing the number of wrongly stemmed words from a number of words and compare results obtained from the algorithm and other algorithms. In the following text, the evaluation is carried out for the RFO Malay stemming approach.

After stop words have been deleted from the Quranic collection, a total of 6,900 distinct words are obtained. The employment of the stemming algorithm reduces the number of these words further to 2,602 distinct roots, or 37.7% for RFO stemmer; whereas, RAO stemmer reduced the number of words to 2667 distinct roots, or 38.6%.

If the level the level of compression is expressed in terms of the number of reduced words, then RFO stemmer achieves 62.3% compression; and RAO stemmer achieves 61.4% compression. The results for each of the stemmer are shown in Table 6.

Table 6: Compression achieved by the algorithms

Stemmer	Distinct words	Compression
RAO	2667	61.4%
RFO	2602	62.3%

RFO stemmer achieved the higher compression. These results indicate that the new stemmer, RFO is the better Malay language stemmer in term of compression ratio. This figure is much higher than similar compression achieved by other language stemmer, English 26.2% to 50.5%, Slovene 54.7% [10].

The percentages of errors obtained by the stemmer for a total of 2116 distinct words are shown in Table 7. This was obtained from experiment on the first two chapters of the Quran. These errors were considered as invalid stems with regards to their Malay language correctness. The highest percentage of 98.6% success rate for Malay stemmer is achieved by RFO.

Table 7: Percentage of errors obtained by the stemmer

Stemmer	Number of errors	Percentage of errors
RAO	93	4.4%
RFO	30	1.4%

As shown in Table 7, the number of errors obtained by Ahmad's [4] stemmer, RAO is 4.4% errors which is superior to the one obtained by our new algorithm, 1.4%. Hence, we can say that the new algorithm, RFO does improve the performance of Malay stemming.

6. Conclusions

From the experiments performed, it is found that, the order of rules to use is not necessary to follow any order of affixes types. However, let the rules sorted in alphabetical order for the first pass, and for the second pass, sort the rules according to usage frequency of each rule. As conclusion from the development of this new Malay stemming algorithm, the order of the rules to be used in the stemming algorithm need to follow the frequency of the application of the rules. Experiments performed in this research showed that the new approaches in stemming are better than other Malay stemmer as RAO by Ahmad [4]. The following research will involve with experimental retrieval system.

References

[1] J. Savoy, Stemming of French Words on Grammatical Categories, Journal of American Society for Information Science, 44(1), pp. 1-9, 1993.

[2] D. Harman, A Failure Analysis on the Limitations of Suffixing in an Online Environment, Proceeding of 10th International Conference on Research and Development in Information Retrieval, pp. 76-85, 1991.

[3] C.J. van Rijsbergen, Information Retrieval, Butterworths, London, 1979.

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

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

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

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

[8] Sock, Y. T., Cheng, S. O., and Abdullah, N. A. 2000. On Designing an Automated Malaysia Stemmer for the Malay language. Proceedings of the 5th International Workshop Information Retrieval with Asian Languages. 207-208.

[9] Idris, N. 2001. Automated Essay Grading System using Nearest Neigbour Technique in Information Retrieval. University of Malaya. Kuala Lumpur.

[10] M. Lennon, An Evaluation of Some Conflation Algorithms for Information Retrieval, Journal of Information Science, 3, pp. 177-183, 1981.

Table 2: Errors for all 7 tests done on 2 chapters of Quran using RFO stemmer

Word	Actual root	Resulting root	Error type	Test no. where error occurs
beribu-ribu	ribu	ibu-ribu	overstemming	1,2,3,4,5,6,7
mengadap	ngadap	adap	overstemming	3,4,6
perasaan	rasa	asa	overstemming	1,2,3,4,5,6,7
bacakan	baca	bacak	understemming	1,2,3,4,5,6,7
bawakan	bawa	bawak	understemming	1,2,3,4,5,6,7
disukai	suka	sukai	understemming	1,2,3,4,5,6,7
memerangi	perang	memerang	understemming	5,7
mengakui	aku	kaku	understemming	1,2,3,4,5,6,7
mengalami	alam	kalam	understemming	1,2,3,4,5,6,7
mengamalkan	amal	kamal	understemming	1,2,3,4,5,6,7
menganiaya	aniaya	kaniaya	understemming	1,2,3,4,5,6,7
mengesahkan	sah	kesah	understemming	1,2,3,4,5,6,7
mengubah	ubah	kubah	understemming	1,2,3,4,5,6,7
mengubahnya	ubah	gubah	understemming	1,2,3,4,5,6,7
mengulangi	ulang	kulang	understemming	1,2,3,4,5,6,7
mengusir	usir	kusir	understemming	1,2,3,4,5,6,7
merasai	rasa	rasai	understemming	1,2,3,4,5,6,7
pengakuan	aku	kaku	understemming	1,2,3,4,5,6,7
penganiayaan	aniaya	kaniaya	understemming	1,2,3,4,5,6,7
peperangan	perang	perangan	understemming	1,2,3,4,5,6,7
usir-mengusir	usir	usir-kusir	understemming	1,2,3,4,5,6,7
sukai	suka	sukai	unchanged	1,2,3,4,5,6,7
memakan	makan	pakan	spelling exception	1,2,3,4,5,6,7
memakannya	makan	pakan	spelling exception	1,2,3,4,5,6,7
memalingkan	paling	maling	spelling exception	1,2,3,4,5,6,7
memejamkan	pejam	mejam	spelling exception	1,2,3,4,5,6,7
memerangi	perang	merang	spelling exception	1,2,3,4,6
pemimpin	pimpin	mimpin	spelling exception	1,2,3,4,5,6,7
pemimpin-pemimpin	pimpin	mimpin	spelling exception	1,2,3,4,5,6,7
penutupnya	tutup	nutup	spelling exception	1,2,3,4,5,6,7
berilah	beri	ilah	others	1,2,3,4,5,6,7
kemari	mari	kemar	others	2,3,4,5,6,7
seruanku	seru	ruan	others	1,3,4

Table 3: Distribution of unique errors using RFO stemmer

Overstemming	Understemming	Unchanged	Spelling exception	Others
3 (9.1%)	18 (54.5%)	1 (3.0%)	8 (24.2%)	3 (9.1%)

Dr. Muhamad Taufik Abdullah is working as Senior Lecturer, Department of Multimedia, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. His areas of interest include Information Retrieval, Cross-Language Information Retrieval and Multimedia Information System.