Experiments with a Stemming Algorithm for Malay Words

Fatimah Ahmad, Mohammed Yusoff, and Tengku M. T. Sembok*
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Malaysia.
E-mail: tmts@fsmk.ukm.my.

Stemming is used in information retrieval systems to reduce variant word forms to common roots in order to improve retrieval effectiveness. As in other languages, there is a need for an effective stemming algorithm for the indexing and retrieval of Malay documents. The Malay stemming algorithm developed by Othman is studied and new versions proposed to enhance its performance. The improvements relate to the order in which the dictionary is looked-up, the order in which the morphological rules are applied, and the number of rules.

Introduction

Lovins (1968) defines a stemming algorithm as "a computational procedure which reduces all words with the same root (or, if prefixes are left untouched, the same stem) to a common form, usually by stripping each word of its derivational and inflectional suffixes." For example, the words *group, groups, grouped, grouping,* or *subgroups* are reduced to the root *group.* A stemming algorithm has applications in the fields of information retrieval and computational linguistics. In information retrieval, grouping words having the same root under the same stem will increase the success with which documents can be matched against a query (Frakes, 1992; Harman, 1991). In computational linguistics, there is a need to identify linguistically correct roots, since the attached affixes provide information about the grammatical function of a word and sometimes to its meaning (e.g., when considering prefixes) and thus help in the syntactic analysis of a sentence.

Many different kinds of stemming algorithms for English have been proposed, ranging from simple procedures that merely remove plurals, past and present participles, to more sophisticated techniques that encompass all of the important types of morphological variation. Such procedures involve either the removal of the single longest matching suffix or the iterative removal of several suffixes, and may also require the specification of detailed context-sensitive rules if there is not to be a significant error rate (Lennon, Peirce, Tarry, & Willett, 1981; van Rijsbergen, 1979). Other reports have detailed stemmers both for languages with a similar morphological structure to that of English, such as Slovene (Popovic & Willett, 1992) and French (Savoy, 1993), and for languages that are less similar, such as Finnish (Jappinen & Ylilammi, 1986) and Turkish (Oflazer, 1994). However, these languages have in common the fact that variant word forms are created by adding suffixes to a basic stem, and the removal of just the suffixes by a stemmer has thus been found to be sufficient for the purposes of information retrieval (Porter, 1980).

The usage of affixes in English (and similar languages) is far less complex than in languages such as Malay (Sembok, Yusoff, & Ahmad, 1994) and Arabic (Al-Kharashi & Evens, 1994), where the stripping of suffixes alone would not be sufficient for retrieval purposes. Thus, focusing on Malay, the root of the words *makanan* (food), *pemakan* (eater), *dimakan* (being eaten), *pemakanan* (diet), and *termakan* (accidentally eaten) is *makan* (eat), and a simple stemmer that removed only suffixes would yield five different stems: *makan, pemakan, dimakan, pemakan,* and *termakan.* It is clear that it is not possible to stem Malay text effectively without considering the removal of prefixes as well as suffixes.

To our knowledge, there is only one existing Malay stemming algorithm that has been developed and tested for applications in information retrieval. This algorithm, which was developed by Othman (1993) as an M.Sc. dissertation project, uses 121 morphological rules, which are arranged and applied to an input word in alphabetical order, and a dictionary of Malay words that is derived from a very large Malaysian dictionary, the *Kamus De-*

* To whom all correspondence should be addressed.
Received April 18, 1995; revised September 13, 1995; accepted September 13, 1995.

wan (DBP, 1991). The dictionary plays an important role in the morphological analysis of Malay: Stemming experiments with and without a dictionary have been reported by Sembok et al. (1994), who show that high error rates and some incomprehensible roots are obtained if a dictionary is not used. In this article, we describe several improvements that we have been able to make to Othman's algorithm to make it better able to handle the morphological complexities of Malay text. The next section provides a brief introduction to these complexities, and this is followed by a discussion of Othman's algorithm and its main characteristics. We then describe the modifications we have made to the original algorithm. The various procedures are tested with two very different data sets, and the article concludes with a summary of our major findings.

Malay Morphology

Malay is spoken by more than 150 million people in Southeast Asia, mostly in Malaysia, Indonesia, and Brunei. In Malaysia, there is a government body, the *Dewan Bahasa dan Pustaka* (Language and Literary Council), which is responsible for creating new Malay words and for controlling and handling all root words in the Malay language. Large numbers of derivatives can be produced from these root words using a set of well-defined classes of affixes and derivation rules. As with other languages, the affixes provide clues in helping the syntactical analysis of a sentence.

There are four main classes of affix. The most frequent is a prefix-suffix pair, i.e., a word that contains both a prefix and a suffix, and the least frequent is the infix, with simple prefixes or suffixes exhibiting intermediate frequencies. For example, an analysis of the two data sets which are used in the experiments reported later in this article gives the following frequencies: For the 736 unique words in 10 chapters from the Malay translation of the Quran, there were 195 prefix-suffix pairs, 140 prefixes, 128 suffixes, 0 infixes, and 273 root words; and for the 434 unique words in a set of 10 research abstracts there were 116 prefix-suffix pairs, 58 prefixes, 44 suffixes, 0 infixes, and 216 root words. As another example, the Malay translation of the complete Quran contains a total of 101,538 word occurrences (including counting the duplicates), of which 16,519 are prefix-suffix pairs, 13,156 contain just prefixes, 13,085 contain just suffixes, and only 12 contain infixes.

In Malay, more than one affix can be attached to a word at the same time, e.g., *memper*juang*kannya* (to fight for it), which involves the following affixes: *mem, per, kan,* and *nya.* But there are constraints on how the affixes are combined. For example, it is semantically wrong to used the affixes *me* and *an* to the word *minum* (drink) to form *me*minum*an*. In the following sections, we shall briefly discuss the classes of affixes in the Malay language.

Prefixes

There are many prefixes available in the Malay morphology: Some of the most important or commonly used ones are *di, ke, se, beR, meN, teR, peN,* and *peR*. Prefixes like *di, ke,* and *se* do not change their form when they are combined with a root word, and there are also not any changes in the spellings of the root words that are attached to them. However, prefixes like *beR, meN, teR, peN,* and *peR* will change their forms, specifically the letters written in capitals below, depending on the first letters of the roots that are attached to them. The following examples illustrate the changes that can occur when a prefix is attached to a root:

di + hantar (send) = *di*hantar (sent (by))
ke + hendak (need) = *ke*hendak (requirement)
beR + rehat (rest) = *be*rehat (resting)
beR + ajar (teach) = *bel*ajar (learning)
beR + tenaga (energy) = *ber*tenaga (energetic)
meN + pakai (wear) = *mem*akai (wearing)
meN + fikir (think) = *mem*ikir (thinking)
teR + minum (drink) = *ter*minum (accidentally drank)
teR + kecil (small) = *ter*kecil (smallest)
peN + nyanyi (sing) = *pe*nyanyi (singer)
peN + sakit (sakit) = *pe*sakit (patient)
peN + ikut (follow) = *peng*ikut (follower)
peR + ajar (teach) = *pel*ajar (student)

Many other prefixes are taken from *loan words,* i.e., words that have been borrowed from foreign languages. Examples include *anti, auto, pro, poli, sub, dwi, pra, eka, foto, feno, hetero, hidro, hiper, inter, kilo, makro, mono, multi, neuro, para, super, tele,* and *tuna.*

Suffixes

Examples of commonly-used Malay suffixes are *an, i, kan, nya, lah,* and *kah.* No variations in spelling are involved when these suffixes are attached to root words, as exemplified below:

makan (eat) + an = makan*an* (food)
harga (value) + i = harga*i* (appreciate)
siap (finish) + kan = siap*kan* (to finish)
rumah (house) + nya = rumah*nya* (his/her/its house)
pergi (go) + lah = pergi*lah* (go (imperative))
siapa (who) + kah = siapa*kah* (who (interrogative))

As with prefixes, many suffixes have been borrowed from foreign languages. Examples include *at, ah, in, atik,*

ator, atis, alis, alisme, alistik, et, eum, grafi, ionil, isme, istik, logi, me, onal, oner, or, rologi, tisme, tologi, uddin, and *tualisme.*

Prefix-Suffix Pairs

As noted previously, this is the most common type of morphology and a correspondingly large number of rules are required to encode them. Some of the frequently-used prefix-suffix pairs are *beR-an, beR-kan, di-i, di-kan, ke-an, meN-i, meN-kan, memper-i, memper-kan, peN-an, peR-an,* and *se-nya.* The spelling exceptions for the root words when attached to this type of affix are the same as for the prefix alone. Examples of the usage of prefix-suffix pairs are as follows:

beR + isteri (wife) + kan = *ber*isteri*kan* (married to)
meN + hadiah (present) + kan = *meng*hadiah*kan* (give a present)
peN + lihat (look) + an = *peng*lihat*an* (sight)
ke + sihat (healthy) + an = *ke*sihat*an* (health)
memper + baik (good) + i = *memper*baik*i* (repair)
se + harus (permissible) + nya = *se*harus*nya* (ought)

Infixes

There are just four infixes available in Malay morphology: *el, em, er,* and *in.* The use of infixes is very rare in Malay language and many people treat the resulting derived words as if they are root words. The infix is always placed between the first and second letters of the root word as illustrated below:

tapak (base) (+el+) = t*el*apak (palm of hand/leg)
gentar (afraid) (+em+) = g*em*entar (shiver)
guruh (thunder) (+em+) = g*em*uruh (nervous)
gigi (tooth) (+er+) = g*er*igi (grill)
sambung (join) (+in+) = s*in*ambung (continue)

Spelling Variations and Exceptions

Prefixes and prefix-suffix pairs may give rise to spelling variations and exceptions in the word root, with the precise form of the variation being determined by the first letter of the attached root, as described previously. For example, the prefix *men* is used only with root words beginning with any of the following letters: *c, d, j, s, t, y,* and *z.*

Spelling exceptions can also occur when the first letter of a root word is dropped on the addition of some prefixes to some root words beginning with certain letters. The specific rules for these exceptions are: With *mem* or *pem* drop *f* or *p*, with *meng* or *peng* drop *k*, with *meny* or *peny* drop *s*, and with *men* or *pen* drop *t.* This behavior is illustrated below:

mem + fikir (think) = *mem*ikir (thinking)
mem + pukul (hit) = *mem*ukul (hitting)
meng + karang (compose) = *meng*arang (composing)
men + tuai (harvest) = *men*uai (harvesting)
meny + simpang (junction) = *meny*impang (deviating)

To handle the spelling exceptions as illustrated above, recoding process has to be used in stemming Malay words. For example, after stripping the prefix *mem* from *memikir,* a letter *f* should be attached to *ikir* to form the correct stem *fikir.*

Variations such as this may not apply when the roots are loan words. There are also some special cases of original Malay words where the first letter of the root word is not dropped when the prefix is attached. Examples of this anomalous behavior include:

Loan words:

mem + fail (file) = *mem*fail (filing) instead of *mem*ail
pem + fitnah (slander) = *pem*fitnah (slanderer) instead of *pem*itnah
peng + khianat (betray) = *peng*khianat (betrayer) instead of *peng*ianat
mem + proses (process) = *mem*proses (processor) instead of *mem*roses
pem + protes (protest) = *pem*protes (protester) instead of *pem*rotes
pen + tadbir (administrate) = *pen*tadbir (administrator) instead of *pen*adbir

Local words:

meng + kaji (investigate) = *meng*kaji (investigating) instead of *meng*aji
men + ternak (rear) = *men*ternak (rearing) instead of *men*ernak

From the above discussion on the Malay morphology, it seems that stemming of Malay words is quite a clear-cut activity in the sense that it is always easy to decide what is the correct stem of a word. In English, by contrast, the stem of a word is extremely ill-defined and it may be said that the aim is not to produce the "correct" stem, but to ensure that any two words referring to the same concept are reduced to the same string. Thus, an evaluation method for Malay stemming algorithms can be based on the percentage of words incorrectly stemmed as we have performed in our experiments.

The Basic Algorithm

The basic algorithm that we have used was originally described by Othman (1993). It adopts a rule-based approach that makes updating of the morphological rules easier. But it slows down the stemming process to as much as 10 times slower than Porter's algorithm. The rules used by Othman are given in Appendix A and will be referred to as set-A rules. These rules define prefixes, suffixes, infixes, and prefix-suffix pairs, and are encoded as follows:

(a) Prefix rules format: Prefix+, e.g., ber+
(b) Suffix rules format: +Suffix, e.g., +kan
(c) Infix rules format: +Infix+, e.g., +el+
(d) Prefix-suffix pair rules format: Prefix + suffix, e.g., di + kan

Affixes are removed through the process of matching the affixes in the rules to that of the input word. The general operation of the algorithm is as follows:

Step-1: If there are no more words then stop, otherwise get the next word;
Step-2: If there are no more rules then accept the word as a root word and go to Step-1, otherwise get the next rule;
Step-3: Check the given pattern of the rule with the word: If the system finds a match, apply the rule to the word to get a stem;
Step-4: Check the stem against the dictionary; perform any necessary recoding and recheck the dictionary;
Step-5: If the stem appears in the dictionary, then this stem is the root of the word and go to Step-1, otherwise go to Step-2.

Three characteristics of this algorithm should be noted. Firstly, the algorithm can overstem, i.e., remove affixes from word roots, because the dictionary is not checked until after the first rule has been applied to the word. For example, the word *masalah* (problem) is overstemmed to *masa* (*time*) when *lah* is considered as a suffix. Secondly, the precise mode of operation of the algorithm depends on the order of the rules, despite the fact that it is not clear in what order the rules should be applied to an input word to obtain the correct root. The morphological complexity of the language means that it is difficult, if not impossible, to derive analytically a "perfect" order of application that will give the correct root for all possible words, and the ordering of the rules must thus be based on arbitrary criteria or on a linguistic "intuition" or "analysis." Othman chose to apply the rules for the four affix classes as follows: First the prefix-suffix (ps) rules, then the prefix (pr) rules, then the suffix (su) rules, and finally the infix (in) rules. The order of the rules within each class is arbitrary. The third characteristic is the total number of rules used. Othman used 121 rules in set-A as listed in Appendix A, but it is not clear that this is sufficient for effective retrieval.

Experimental Details

The Rule Sets

The experiments reported here have adopted Othman's general algorithm and set of morphological rules as the basis for further development. We soon found that his set of 121 rules does not cover many of the affixes in Malay. For example, *ke* + *anku* as in *ke*sedih*anku, pe* + *anmu* as in *pe*mergi*anmu, +ullah* as in bait*ullah, +al* as in klinik*al,* and *+si* as in formula*si* are not included in the set.

Accordingly, additional rules were developed by:

- Exhaustive scanning of entry words in a Malay dictionary (DBP, 1991);
- Exhaustive scanning of a book on Malay spelling (DBP, 1987);
- reference to a book on Malay morphology (Karim, Onn, & Musa, 1993);
- reference to the two data sets used for our experiments. These were 10 chapters taken from the Malay translation of the Quran (Hamidy and Fachruddin, 1987) and 10 research abstracts on research done at Universiti Kebangsaan Malaysia (Sharifah Mastura, Ungku Maimunah, & Ramli, 1989).

These analyses resulted in the development of two new rule sets. The first set (set-B) contains the 432 rules detailed in Appendix B while the second (set-C) contains the 561 rules detailed in Appendix C. The two sets differ only in that set-C includes rules that cater to the modern Malay derivatives that were identified in our second test data set. Both sets include all the rules in the set-A as listed in Appendix A.

The Experiments

Three main groups of experiments were carried out, as follows:

(1) To determine the effect of checking the input word first against the dictionary.
(2) To find the best order in which rules are applied in the stemming process. The experiments are performed on two data sets, i.e., 10 chapters of the Quran translation and 10 research abstracts, to determine which group of rules to trigger first to produce the minimum number of errors. As mentioned

TABLE 1. Effect of initial dictionary check on numbers of errors. Three chapters from Quran. Six different rule orderings.

	No initial dictionary check[b]						Initial dictionary check[b]					
Rules from set-B[a]	1	2	3	4	5	6	1	2	3	4	5	6
660+	19	19	32	36	22	35	6	6	19	24	11	24
420~	15	15	18	21	17	20	6	6	9	13	10	13
371*	13	13	15	16	13	15	6	6	8	10	8	10

[a] Total no. of words stemmed.
[b] Test no.
+, Total no. of word occurrences in all the chapters.
~, Total no. of unique words counted within each chapter.
*, Total no. of unique words in all the chapters combined together.

earlier the rules are grouped into prefix (pr), suffix (su), prefix-suffix pair (ps), and infix (in) rules. The order of rules within the four classes of rules is arbitrary and order-dependent. We have tested all possible orderings of the four classes, with the exception that the infix rules are always tested last, owing to their rarity in Malay. The order of application of the rules is thus:

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.

(3) To evaluate the relative merits of the sets of rules that are defined in Appendices A–C.

Experimental Results and Discussion

Initial Checking of the Dictionary

These experiments were carried out to determine the effect of checking each input word against a dictionary of root words before it was processed by the stemmer, which in this case was based on the 432 rules listed in Appendix B. The initial checking against the dictionary will avoid doing stemming on words which are already root words. The source data used here was three chapters from the Malay version of the Quran.

The first set of six runs used the basic Othman algorithm, in which the dictionary is not first checked before the application of the stemmer, with each of the six rule-orderings defined in the previous section. The only difference in the second set of six runs was that an initial check was carried out. The dictionary that was used was a version of the *Kamus Dewan* (DBP, 1991) that had been manually modified to contain a total of 22,393 word roots.

The performance of the various procedures was measured by the numbers of words that were stemmed incorrectly as compared to the stems produced manually. The numbers of words incorrectly stemmed are obtained by running the various procedures on three different collections of words from the data sets as follows: All word occurrences in each chapter/abstract are stemmed; all unique words within each chapter/abstract are stemmed; and only unique words among all the chapters/abstracts are stemmed. The total numbers of words stemmed for the three collections are marked with the symbols +, ~, and *, respectively, in Tables 1–3 and 6. Table 1 shows the results obtained from the two groups of experiments, and it will be seen that there is a significantly lower error rate if an initial dictionary check is carried out prior to the invocation of the stemmer. It

TABLE 2. Effect of different rule orderings on numbers of errors. Ten chapters from Quran. Six different rule orderings and other approaches.

	Test no.						Others		
Rules from set-B[a]	1	2	3	4	5	6	Shortest	Longest	Othman[b]
1548+	13	17	36	43	24	47	32	255	142
1053~	13	15	20	26	21	28	30	170	116
736*	12	12	13	16	15	16	24	105	96

[a] Total no. of words stemmed.
[b] Rules from set-A are used in Othman's original algorithm.

TABLE 3. Effect of different rule orderings on numbers of errors. Ten research abstracts. Six different rule orderings and other approaches.

	Test no.						Others		
Rules from set-B[a]	1	2	3	4	5	6	Shortest	Longest	Othman[b]
757+	14	17	11	17	23	20	23	132	61
504~	12	13	9	14	18	15	16	90	38
434*	11	12	8	13	17	14	16	76	33

[a] Total no. of words stemmed.
[b] Rules from set-A are used in Othman's original algorithm.

will also be seen that the best results are obtained with Test-1 and Test-2, i.e., when the prefix rules are used first.

Order of Application of the Rule Sets

We have studied the order of application in more detail using the set-B rules with our two principal data sets, viz 10 chapters of the Quran and 10 research abstracts. For comparison, we have also run Othman's original algorithm and two versions of the set-B rules, which we shall refer to as *longest match* and *shortest match*. In the longest-match approach, we simply remove the single longest affix that can be mapped to the input word, thus yielding the shortest root possible from the removal of a single affix (and conversely for the shortest-match approach).

The results obtained are listed in Tables 2 and 3. It will be seen from these tables that there are only minor differences between the six class orderings that can be used with the set-B rules. However, there are differences between the Quran and abstracts data sets and it is not possible to say that any one of the approaches is unequivocally the best. There is, however, no doubt that the other three approaches, i.e., the original algorithm and the shortest (or longest) match algorithms, yield many more incorrect word roots.

The types of error produced by Tests 1–6 are shown in Tables 4 and 5, which are for the Quran and research-abstracts data sets, respectively. The errors have been

TABLE 4. Errors from experiments using the Quran data set.

Word	Actual root	Resulting root	Error type	Test no. where error occurs
kurangkan	kurang	rang	Overstemming	1, 3, 4
memakan	makan	mak	Overstemming	3, 4, 6
sebabnya	sebab	bab	Overstemming	1, 3, 4
berduri	duri	dur	Overstemming	3, 4, 6
peringatan	ingat	peringat	Understemming	2, 5, 6
sepenuh	penuh	sepenuh	Unchanged	1, 2, 3, 4, 5, 6
perangan	perang	perangan	Unchanged	1, 2, 3, 4, 5, 6
berbuah	buah	berbuah	Unchanged	1, 2, 3, 4, 5, 6
sebelah	belah	sebelah	Unchanged	1, 2, 3, 4, 5, 6
sukai	suka	sukai	Unchanged	1, 2, 3, 4, 5, 6
mengandung	kandung	gandung	Spelling exception	1, 2, 3, 4, 5, 6
berilah	beri	ilah	Others	1, 2, 3
berikanlah	beri	ikan	Others	1, 3, 4
beriman	iman	rim	Others	3, 4, 6
mencari	cari	pencar	Others	4, 5, 6
keburukan	buruk	keburu	Others	2, 5, 6
melihat	lihat	pelih	Others	4, 5, 6
sekali	kali	sekal	Others	4, 5, 6
melengah	lengah	meleng	Others	4, 5, 6
memulai	mula	pulai	Others	1, 2, 5
kesenangan	senang	tangan	Others	1, 2, 5
menurunkan	turun	penurun	Others	2, 5, 6

TABLE 5. Errors from experiments using the research abstracts data set.

Word	Actual root	Resulting root	Error type	Test no. where error occurs
kemudiannya	kemudian	mudi	Overstemming	1, 3, 4
berkesan	kesan	kes	Overstemming	3, 4, 6
diredai	reda	redai	Understemming	1, 2, 5
pengesahan	sah	sahan	Understemming	1, 2, 5
kedudukan	duduk	keduduk	Understemming	2, 5, 6
menerusi	terus	terusi	Understemming	1, 2, 5
ketimbulan	timbul	ketimbul	Understemming	2, 5, 6
formulasi	formula	formulasi	Unchanged	1, 2, 3, 4, 5, 6
membangun	bangun	membangun	Unchanged	1, 2, 3, 4, 5, 6
realisasi	realis	realisasi	Unchanged	1, 2, 3, 4, 5, 6
klinikal	klinki	klinikal	Unchanged	1, 2, 3, 4, 5, 6
mengandungi	kandung	gandung	Spelling exception	1, 2, 3, 4, 5, 6
pengarang	karang	garang	Spelling exception	1, 2, 3, 4, 5, 6
lelaki	laki	lelak	Others	4, 5, 6
melihat	lihat	pelih	Others	4, 5, 6
sekali	kali	sekal	Others	4, 5, 6
kesenangan	senang	tangan	Others	1, 2, 5
sebuah	buah	sebu	Others	4, 5, 6
terikat	ikat	terik	Others	4, 5, 6

classified into five classes: *Overstemming, understemming, unchanged, spelling exception,* and *others.* The first three have been identified by Savoy (1993) and the other two are introduced to cater specifically to errors encountered in stemming of Malay words. *Overstemming* occurs when more characters have been removed from the input word than necessary. On the other hand, *understemming* occurs when too few characters have been removed. *Unchanged* is in fact a special case of understemming. It occurs when no characters have been removed when some should have been removed in order to get the correct root. *Spelling exception* occurs when the first letter of the stem obtained is not correctly recoded after the prefix has been removed. Other types of error are classified as *others.*

The largest class of errors is that called *others.* These are mostly due to the order in which the rules are applied when a word is stemmed. The second most common error type is *unchanged.* These errors mainly occur with words that can either be a root word themselves or be a derivative of some other word, although such an error can also occur if there are some non-root words in the dictionary of (supposedly) root words. The words *formulasi, realisasi,* and *klinikal* remained unchanged in the research-abstracts data set since set-B does not include any rules for the derivation of modern Malay words from foreign words that use the suffixes *si, asi,* and *al.* The effect of adding such rules was explored in the final set of experiments.

Use of an Extended Set of Rules

The final experiments used the extended set of rules, set-C, and applied them to the Quran and research-abstracts data sets, as shown in Table 6. For comparison, we have also included again the results obtained with

TABLE 6. Numbers of errors with rule set-C compared to rule set-B.

	Rules			Rules	
Quran data set[a]	Set-B	Set-C	Research abstracts[a]	Set-B	Set-C
1548+	13	19	757+	11	6
1053~	13	17	504~	9	5
736*	12	15	434*	8	5

[a] Total no. of words stemmed.

the smaller set of rules, set-B. In both cases, we have used that ordering which gave the best results in the experiments described previously, i.e., Test-1 for the Quran data set and Test-3 for the research-abstracts data set.

Set-C includes rules that are specifically designed to encompass modern Malay words. The results in Table 6 are thus precisely those that might have been expected since performance is improved, i.e., there is a smaller number of errors, when the modern, research-abstracts data set is processed by this set of rules; however, set-B gives the better level of performance with the classical Quran data set. Thus, three of the Quran words (*biarkan, hatimu,* and *warnanya*) were wrongly stemmed (to *arkan, timu,* and *nya,* respectively) owing to the inclusion of rules for the prefixes *bi, ha,* and *warna* in set-C. On the other hand, the addition of rules for the suffixes *si, asi,* and *al* in set-C means that the words *formulasi, realisasi,* and *klinikal* are now correctly stemmed to *formula, realis,* and *klinik,* respectively.

Conclusions

We believe that the identification of the correct root for each word in a text is vital for the automatic indexing of Malay documents, and in this article, we have evaluated several approaches to the identification of such roots. Our experiments have demonstrated clearly that there are several, simple modifications that can be made to Othman's stemmer that will significantly increase its ability to stem Malay words correctly. Thus, many errors can be eliminated by checking a dictionary before any of the rules are applied and others can be eliminated by expanding the number of rules. However, we have obtained different results with the ancient and modern data sets, with the smaller set of rules, set-B, giving the better performance with the Quran data set and the larger set of rules, set-C, giving the better performance with the research-abstracts data set.

Our experiments have shown that the new versions of the algorithm perform much better than the original version. Could it be improved further? Our analysis suggests that most of the remaining errors are due to the precise order in which the rules are applied within each of the four classes of rules, and we are currently considering ways in which this ordering can be best optimized.

Nevertheless, in an information retrieval context, the way to really evaluate a Malay stemming algorithm is to carry out some information retrieval performance evaluation, with recall-precision analysis, on one or more test collections in the Malay language. This will logically be our next step to work on.

Acknowledgments

We thank Professor P. Willett, Department of Information Studies, University of Sheffield, for reading initial drafts of this article and for his invaluable comments and suggestions. We thank the Association of Commonwealth Universities for the Fellowship grant to the last author, Tengku M. T. Sembok, to spend his sabbatical leave at the University of Sheffield where this article was completed, and to the Malaysian Ministry of Science and Technology for the IRPA grant (02-07-03-024) which provided the computer facility for this research.

Appendix A

Rules of set-A (Othman's collection).

+an	dwi+an	mono+
+annya	eka+	panca+
+at	juru+	pe+
+el+	ke+	pe+an
+em+	ke+an	pe+annya
+er+	ke+annya	pe+wan
+grafi	ke+nya	pe+wati
+i	keber+an	pel+
+iah	kepel+an	pel+an
+ilah	maha+	pem+
+in	me+	pem+an
+is	me+i	pembel+an
+isasi	me+inya	pen+
+isme	me+kan	pen+an
+kan	mem+	pen+annya
+lah	mem+i	peng+
+logi	mem+kan	peng+an
+man	membe+	penge+
+nya	mempe+kan	penge+an
+wan	mempel+kan	peny+
+wi	memper+	peny+an
anti+	memper+i	per+
antike+an	memper+inya	per+an
be+	memper+kan	per+annya
bel+	men+	poli+
ber+	men+i	pra+
ber+an	men+inya	pro+
ber+kan	men+kan	se+
berke+an	meng+	se+an
berpen+	meng+i	se+kan
dasa+	meng+inya	se+nya
di+	meng+kan	seber+
di+i	menge+	sub+
di+kan	menge+i	sub+an
di+nya	menge+kan	tata+
dike+kan	menter+kannya	ter+
dike+kannya	menterke+kannya	ter+lah
diper+an	meny+	ter+nya
diper+annya	meny+i	terper+
diper+kan	meny+kan	tri+
dwi+		

Appendix B

Rules of set-B (cater to words in the Quran).*

+anda	be+kan	me+bobokkan	men+ani	pen+gunaan
+anku	be+lah	me+gandakan	men+ankan	pen+kan
+anmu	de+	me+kanmu	men+balikkan	pen+ku
+ah	di+inya	me+kannya	men+baurkan	pen+mu
+cita	di+kankah	me+lah	men+biakkan	pen+nya
+in+	di+kannya	me+lecuhkan	men+ertikan	pen+pastian
+iah	di+lah	me+mu	men+fahamkan	peny+nya
+ial	dike+i	me+negarakan	men+gunakan	per+anku
+iat	dike+inya	me+nya	men+hitamkan	per+anmu
+ina	dimeng+	me+telentangkan	men+kanmu	per+i
+inya	dipeng+kan	mem+hanguskan	men+kannya	per+ilah
+kah	diper+	mem+inya	men+lah	per+kan
+kali	diper+i	mem+kanmu	men+luaskan	per+kanlah
+kankah	diper+inya	mem+kannya	men+mu	per+lah
+kanlah	diper+kannya	mem+mu	men+nya	per+nya
+kannya	diper+nya	mem+nya	men+padukan	per+wan
+kata	dise+apakan	mem+wankan	men+rasulkan	perike+an
+kaukah	dise+kan	member+kan	men+ratakan	permaha+an
+ku	diter+kan	meme+kan	men+tafsirkan	perse+an
+loka	ge+	memer+	menganak+kan	perse+annya
+mana	ge+an	memer+kan	meny+kannya	pra+an
+mu	je+	mempe+	meny+nya	re+
+mulah	juru+is	mempel+i	menyalah+kan	re+an
+neka	ke+anku	mempen+kan	menyatu+kan	sepeng+
+nyakah	ke+anmu	mempeng+kan	menye+	sepeng+an
+nyaku	ke+biasaan	memper+belikan	menye+i	seper+an
+pun	ke+i	memper+kannya	menye+kan	seper+nya
+sanya	ke+ilah	memper+nya	mer+	sepen+
+tah	ke+kan	memper+padukan	para+	sepen+an
+tari	ke+kanlah	memperanak+kan	pe+anku	sepen+nya
+uddin	ke+ragaman	memperke+kan	pe+anmu	sede+
+ullah	ke+rataan	memperse+i	pe+kan	seke+
+wati	ke+wanan	memperse+kan	pe+mu	seke+an
+wiah	kedwi+an	mene+	pe+nya	seke+nya
al+	kejuru+an	mengem+	pe+wati	sese+
antipen+	keke+an	mengem+i	pel+i	sepe+
ber+belakangkan	kemaha+an	mengke+i	pem+anku	sepe+an
ber+i	keme+an	mengke+kan	pem+anmu	sepe+anku
ber+kah	kemen+an	mense+kan	pem+annya	sepe+anmu
ber+lah	kepe+an	menter+i	pem+ku	sepe+annya
ber+nya	kepen+an	menter+kan	pem+mu	sepe+nya
berke+	kepeng+an	menge+inya	pem+nya	se+annya
berke+anlah	keper+an	meng+kannya	pember+an	se+ku
berke+i	kese+	meng+anginkan	pemer+	se+mu
berle+	kese+an	meng+bahasakan	pemer+an	se+pun
bermaha+	kese+annya	meng+isasi	penge+anku	sem+
berme+	kesu+an	meng+isasikan	penge+anmu	sem+an
berpe+	ketak+an	meng+isasikannya	penge+annya	te+
berpe+an	ketata+an	meng+jalinkan	pense+an	tata+an
berpel+an	keter+an	meng+kabulkan	penye+	terbe+
berpem+an	keter+annya	meng+lah	penye+an	terke+
berpen+an	ketidak+an	meng+leburkan	perse+	terpe+
berpeng+	ketidakse+an	meng+mu	peng+anku	terpel+
berpeng+an	ku+	meng+mukakan	peng+anmu	terse+
berpenge+an	ku+kan	meng+nikahkan	peng+annya	teper+
berper+	ku+kankah	meng+nya	peng+jalinan	ter+i
berper+an	ku+kanlah	meng+panaskan	peng+ku	ter+kan
berse+	le+	meng+putihkan	peng+mu	ter+an
berse+an	me+an	meng+sertakan	peng+nya	te+an
berse+kan	me+balikkan	men+adukkan	pen+anku	te+kan
berseke+an	me+belahkan	men+ajarkan	pen+anmu	tele+
berte+an	me+belitkan	men+alukan	pen+annya	warga+
be+an				

* Rules of set-A are also included in this set but not listed here.

Appendix C

Rules of set-C (cater to modern Melay words).*

+a	+tual	men+biakkan
+andus	+tualisme	men+ertikan
+aris	+us	men+fahamkan
+asi	auto+	men+gunakan
+atik	bersi+	men+hitamkan
+atis	bersi+an	meta+
+ator	bersimaha+	meta+al
+al	bi+	mikro+
+alis	ce+	mili+
+alisasi	dasa+	mono+
+alisme	dwi+	mono+i
+alistik	dwi+an	multi+
+et	dwi+wan	multi+an
+eum	eka+	neo+
+if	feno+	neuro+
+il	feno+an	neuro+an
+ionil	fito+	oksi+
+isir	foto+	orto+
+isma	ha+	paleo+
+isme	heksa+	penta+
+istik	hidro+	peri+
+itas	hiper+	perike+an
+iter	homo+	permaha+an
+itet	infra+	petro+
+iti	inter+	pleo+
+logi	intra+	poli+
+me	intra+an	pro+
+nita	iso+	si+
+onal	ka+	sub+
+oner	kaji+	sub+an
+or	kaji+an	super+
+rologi	kilo+	super+an
+san	manca+	supra+
+sasi	makro+	supra+an
+si	memono+	te+an
+tal	menganak+kan	te+kan
+tari	men+adukkan	tele+grafi
+tasi	men+ajarkan	tri+
+tik	men+alukan	tri+an
+tis	men+ani	tripra+
+tisme	men+ankan	tuna+
+tologi	men+balikkan	ultra+
+tor	men+baurkan	warna+

* Rules of set-B are also included in this set but not listed here.

References

Al-Kharashi, I. A., & Evens, M. W. (1994). Comparing words, stems, and roots as index terms in an Arabic information retrieval system. *Journal of the American Society for Information Science, 45,* 548–560.

DBP (1987). *Daftar Ejaan Rumi Bahasa dan Pustaka.* Kuala Lumpur, Malaysia: Dewan Bahasa dan Pustaka.

DBP (1991). *Kamus Dewan* (Edisi Baru). Kuala Lumpur, Malaysia: Dewan Bahasa dan Pustaka.

Frakes, W. B. (1992). Stemming algorithms. In W. B. Frakes, R. Baeza-Yates (Ed.), *Information retrieval: Data structures & algorithms* (pp. 131–160). Englewood Cliffs, NJ: Prentice-Hall.

Hamidy, Z. H., & Fachruddin, Hs. (1987). *Tafsir Quran.* Klang, Malaysia: Klang Book Centre.

Harman, D. (1991). How effective is suffixing? *Journal of the American Society for Information Science, 42,* 7–15.

Jappinen, H., & Ylilammi, M. (1986). Associative model of morphological analysis: An empirical inquiry. *Computational Linguistics, 12,* 257–272.

Karim, N. S., Onn, F. M., & Musa, H. (1993). *Tatabahasa Dewan.* Kuala Lumpur, Malaysia: Dewan Bahasa dan Pustaka.

Lennon, M., Peirce, D. S., Tarry, B. D., & Willett, P. (1981). An evaluation of some conflation algorithms for information retrieval. *Journal of Information Science, 3,* 177–183.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics, 11,* 22–31.

Oflazer, K. (1994). Two-level description of Turkish morphology. *Literary and Linguistic Computing, 9,* 175–198.

Othman, A. (1993). *Pengakar Perkataan Melayu untuk Sistem Capaian Dokumen.* Unpublished master's thesis, Universiti Kebangsaan Malaysia, Bangi, Malaysia.

Popovic, M., & Willett, P. (1992). The effectiveness of stemming for natural-language access to Slovene textual data. *Journal of the American Society for Information Science, 43,* 384–390.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14,* 130–137.

Savoy, J. (1993). Stemming of French words based on grammatical categories. *Journal of the American Society for Information Science, 44,* 1–9.

Sembok, T. M. T., Yusoff, M., & Ahmad, F. (1994). A Malay stemming algorithm for information retrieval. *Proceedings of the 4th International Conference and Exhibition on Multi-Lingual Computing.* London, (pp. 5.1.2.1–5.1.2.10). Cambridge, England: University of Cambridge.

Sharifah Mastura, S. A., Ungku Maimunah, M. T., & Ramli, M. S. (1989). *Research abstracts: UKM 1980–1989.* Bangi, Malaysia: Universiti Kebangsaan Malaysia.

van Rijsbergen, C. J. (1979). *Information retrieval.* London: Butterworths.