Enhanced Affixation Word Stemmer with Stemming Error Reducer to Solve Affixation Stemming Errors

Mohamad Nizam Kassim[1], Mohd Aizaini Maarof[2], Anazida Zainal[3] Amirudin Abdul Wahab[4]
[1,4]*Strategic Research, CyberSecurity Malaysia,*
The Mines Resort City, 43300 Seri Kembangan, Malaysia.
[1,2,3]*Faculty of Computing, Universiti Teknologi Malaysia,*
81310 Skudai Johore, Malaysia.
nizam@cybersecurity.my

***Abstract*— Word stemming algorithm (or word stemmer) is an important preprocessing component in the information retrieval and text categorization that aims to reduce derived words to their respective root words. Most of the existing Malay word stemmers adopt rule-based affixes removal method and dictionary lookup to stem affixation words. Despite of many stemming approaches have been proposed in the past research, the existing Malay word stemmers still suffer from affixation stemming errors due to the complexity of Malay morphology. These stemming errors can be classified into over stemming, under stemming, unstem, and special variations and exceptions. Hence this paper presents the enhanced affixation word stemmer that aims to solve these stemming errors. This paper also examined the root causes of these stemming errors in the existing Malay stemmers. The experimental results indicate that the enhanced word stemmerable to stem prefixation, suffixation, confixation and infixation wordswith better stemming accuracy by using enhanced Rule Application Order and Stemming Errors Reducer.**

***Index Terms*— Malay word stemmer; Word stemming algorithm; Word stemmer; Stemming error; Rules application order.**

I. INTRODUCTION

In general, word stemming is a morphology process of natural language to reduce the derived words to their respective root words. It has been widely applied in various fields such as information retrieval, text categorization, text mining and machine translation [3][9][16]. The program used to perform word stemming process is called word stemming algorithm or word stemmer. Word stemmer is developed based on specific morphology structures of natural language such as grammatical structure, language characters and syntactical of that natural language [6][11][14]. Every natural language contains many different word variants of their respective root words [1-2]. For instance, English words '*connects*', '*connected*' and '*connection*' are derived from its root word of '*connect*'. However, Malay language has very complex morphology rules than English language where Malay words of '*menyambung*', '*sambungan*', '*penyambungan*' and '*sinambung*' are derived from its root word of '*sambung*' in which removing suffixes alone are not sufficient [1-2][13]. Therefore, past researchers have developed many word stemmers for Malay language with various stemming approaches that include the best order of prefixation, suffixation, confixation and infixation [1-2][4-5][10][13], selection of longest/shortest match to identify the affixes [1-2][7], the order of dictionary lookup [2][9][13] and dictionary entries [4][9][13]. Regardless of various word stemming approaches that have been adopted in the past research, the existing Malay word stemmers still suffer from affixation stemming errors [1-3][7][10][13][16]. The root causes of these stemming errors are due to affixes removal method, root words that are similar to affixation words and multiple spelling variation and exceptions rules in prefixation and confixation [1-2][7]. Therefore it is desirable to select the best stemming approach to suit with Malay morphology that demonstrates better stemming accuracy to address these affixation stemming errors. Hence, this paper proposes an enhanced Rules Application Order approach with Stemming Error Reducer to address these stemming errors. This paper is organized into five subsequent sections. Section 2 discusses related works on the existing Malay word stemmers. Section 3 describes the root causes of affixation stemming errors in the existing Malay word stemmer. Section 4 discusses the proposed affixation word stemmer. Section 5 discusses the experimental results and discussion where Malay online news articles have been used to evaluate the proposed stemmer. Finally, Section 6 concludes this paper with a summary.

II. THE EXISTING MALAY WORD STEMMERS

Lovins (1968), Dawson (1974) and Porter (1980) contributed to the earliest research on word stemmers for English language that have sparked other word stemmer research in various natural languages such as English, Latin, Dutch and Arabic [1-2][7]. These research have led to active research on subsequent word stemmers in the information retrieval [2][9][12]. The most prominent word stemmer in the literature is Porter Stemmer [1-2][7]. It has become a standard word stemmer for English language and the same stemming approach has been adopted to other natural languages that are Romance (French, Italian, Portuguese and Spanish), Germanic (Dutch and German), Scandinavian languages (Danish, Norwegian and Swedish), Finnish and Russian [12]. On other hand, Othman (1993), Sembok et al. (1994), Ahmad et al. (1996) pioneered the earliest research on word stemmers for

Malay language. These research gave research direction on subsequent word stemmers with various stemming approaches [1-5][7-10][12-13][15][16]. It can be deduced that there were two research directions in word stemmers for Malay language. The first research direction is to improve rule-based word stemmer developed by Othman (1993). Rule-based stemming approach contains sets of affixation word stemming rules arranged in alphabetical order and dictionary lookup to check the word is not root word before word stemming [2][10]. However Sembok et al. (1994) has highlighted the weaknesses using rule-based stemming approach [13]. Therefore the rule application order stemming approach was introduced by rearranging affixation word stemming rules in morphological order and dictionary lookup was used before and after word stemming rules in order to check whether the word is a root word [2][7][15]. Then, the rule frequency order stemming approach was introduced to con-sider only most frequent affixes and root word dictionary lookup was also used before and after word stemming in order to check whether the word is a root word [1][4][5]. Finally, modified the rule frequency order was introduced to consider only most frequent affixes and having two types of dictionaries called background knowledge [9]. These rule-based stemming approaches were introduced to improve affixation word stemming accuracy. The second research direction is by adopting other stemming approaches such as modified Porter Stemmer [12] and Boolean extraction method [16]. However, these stemming approaches are not popular among past researchers due to the complexity of Malay morphology to find the correct root wordsfrom similar derived words such as *menyanyi* (singing) →*nyanyi* (sing), *menyonteng* (scribbling on) →*conteng* (scribble) and *menyapu* (sweeping) →*sapu* (sweep). Moreover, there are other factors that affect the performance of these stemming approaches such as number of affixation stemming rules to stem prefixation, suffixation, confixation and infixation, the numbers of word entriesfor dictionary lookup,and spelling variations and exceptions in word stemming [1-5][7-10][12-13][15-16].

III. Affixation Words And Affixation Stemming Errors

In Malay morphology, affixation words may contain prefixes (at the beginning of the derived word), suffixes (at the ending of the derived word), confixes (at the beginning and ending of the derived word) and infixes (at the middle of the derived word) attached to the root words as shown in Table 1.

Table 1
Combination of Root Words with Affixes, Clitics and Particles

Affixation	Combination of Root Words with Affixes, Clitics and Particles
Prefixation	1. prefix+ root word: [*pembaca*(reader)] 2. multiple prefixes + root word: [*mempersenda* (joking)] 3. prefix + root word + particle: [*bergembiralah* (have fun)]
Suffixation	1. root word + suffix: [makanan (food)] 2. root word+suffix+enclitic/particles [makananmu (his/her food)]
Confixation	1. prefix + root word + suffix: [*pertanian* (agriculture)] 2. multiple prefixes+root word+suffix: [memperjudikan (gamble)] 3. prefix + root word + suffix + enclitic/particles [*pelajarannya* (his/her study)] 4. multiple prefixes + root word + suffix + enclitic/particles:[*mempertemukannya* (his/her findings)] 5. proclitic + root word + suffix: *kunantikan* (I'm waiting)]:(ku)+(nanti)
Infixation	root word with infix: [*telunjuk* (pointing fingers) →*tunjuk*]

In some instances, proclitic, enclitic and particles must also be removed if attached to affixation. There is additional morphological rules called special variations and ex-ceptions in Malay language in the selected prefixes:to remove *men+*, *pen+* and re-place with letter t [*menolong* (helping) →*tolong* (help)], to remove *mem+*, *pem+* and replace with letter f,p[*memilih* (choosing) →*pilih* (choose) and *memikir* (thinking) →*fikir* (think)], to remove meng+, peng+ and replace with letter k [*mengayuh* (ped-dling/cycling) →*kayuh* (peddle/cycle)] and to remove *meny+*, *peny+* and replace with letter s [*menyebut* (saying) →sebut (say)]. Based on the existing literature on Malay word stemmers, past researchers have classified affixation stemming errors in terms of overstemming, understemming, unstem, and spelling variations and exceptions [1-3][7][10][13][16]. However, this paper redefines these stemming errors based on their respective root causes as follows:

i. Affixation Stemming Errors Type I

Applying affixation word stemming rules against root words that are similar to affixation words e.g. prefixation word stemming: *terjemah* (translate) →*terjemah* (not jemah),suffixation word stemming: *makan* (eat) →*makan* (not mak) and confixation word stemming: *menterai* (signing) →*mente-rai* (not tera).

ii. Affixation Stemming Errors Type II

Applying confixation word stemming rules against prefixation words e.g. *berpedoman* (be guided) →*pedoman* (not pedom), *diberi* (given by) →*beri* (not beri) and *memakan* (eating) →*makan* (not mak).

iii. Affixation Stemming Errors Type III

Applying confixation word stemming rules against suffixation words e.g. *bersihkan* (cleaning up) →*bersih* (not sih), *kedainya* (his/her shop) →*kedai* (not da) and *pekikan* (shout) →*pekik* (not ki).

iv. Affixation Stemming Errors Type IV

Applying incorrect affixes removal selection against prefixation word stemming e.g.prefixes (be+ and ber+): *berasa* (to feel) →*rasa* (not asa) and *beranak* (to give birth) →*anak* (not ranak).

v. Affixation Stemming Errors Type V

Applying incorrect affixes removal selection against suffixation word stemming e.g. suffixes (+anlah, +kanlah, and +lah):makanlah (to eat) →makan (not ma, mak) and biarkanlah (let it be) →biar (not biark, biarkan).

vi. Affixation Stemming Errors Type VI

Applying incorrect affixes removal selection against prefixation and confixation words e.g. *menge+*, *meng+*, *meny+*, *men+*, *me+*:*mengerikan* (too eerie) →*ngeri* (*not eri, keri, ri*); *mengenalkan* (to introduce) →*kenal* (*not ngenal, enal, nal*); *mengeratkan* (to tighten) →*erat* (*not ngerat, kerat, rat*) and *mengetinkan* (*to can up*) →*tin* (*not ngetin, ketin, etin*), *menyanyi* (to sing) →*nyanyi* (*not sanyi, yanyi*), *menyapu* (to sweep) →*sapu* (not nyapu, yapu).

vii. Affixation Stemming Errors Type VII

Applying incorrect affixes removal selection against multiple affixes in prefixation, suffixation, confixation and infixationwords e.g. *pemergianmu* (your departure) →*pergi* (*not mergi, erg*), *terangkanlah* (please explain) →*terang* (*not terangk, terangkan*), *kesinambungan* (continuity) →*sambung* (not sinambung).

viii. Affixation Stemming Errors Type VIII

Insufficient or no stemming rules to stem prefixation, suffixation, confixation and infixation e.g. *mengenalkan* (to introduce) →*kenal* (*not mengenalkan, ngenalk, ngenal, genal, enal*).

Therefore, it is highly desirable to develop new stemming approach to address these stemming errors in the existing Malay word stemmers. The stemming approach ought to address applying word stemming rules against the correct word patterns (Affixation Stemming Errors Type I, II and III), ensuring affixes removal selection against the correct affixes in affixation words (Affixation Stemming Errors Type IV, V, VI and VII) and developing sufficient word stemming rules based on morphological structures in Malay language (Affixation Stemming Errors Type VIII).

IV. The Proposed Affixation Word Stemmer

Figure 1 describes the proposed affixation word stemmer that uses enhanced rules application order to stem prefixation, suffixation, confixation and infixation words. This proposed affixation word stemmer was developed using Perl Programming v5.5 on MacBook Pro OS X El Capitan with 2.8 GHz Intel Core i7 processor and 8 GB 1600 MHz DDR3 memory. There are four different affixation stemming rules were developed to match and remove affixes based on Malay morphology book by Abdullah Hassan[6] i.e. 57 prefixation word stemming rules, 42 suffixation word stemming rules, 270 confixation word stemming rules and 1 table lookup rule for infixation word stemming.The best of order for affixation word stemming is selected based on experiment by Ahmad et al. (1996): *infixation* → *confixation*→ *prefixation*→ *suffixation*. The number of affixation stemming rules in the proposed affixation word stemmer is 370 rules which is lessthanSembok's stemmer (432 rules) to stem prefixation, suffixation, confixation and infixation words[1][2][3].To address the existing affixation stemming errors(Affixation Stemming Errors Type I, II, III, IV, V, VI, VII and VIII), the proposed word stemmer also uses three different types of dictionaries called Stemming Errors Reducer. Stemming Errors Reducer comprises of *lookup dictionary* that contains infixation words(41 words), *root word dictionaries* that contain root words similar to prefixation words (2403 words), suffixation words (1448 words) and confixation words (87 words) and *derivative dictionaries* that contain multiple rules of prefixation words (579 words), suffixation words (258 words) and confixation words (1155 words). These dictionaries are used to stem infixation (*lookup dictionary*), to address stemming errors due to root words that are similar to prefixation, suffixation and confixation words (*root word dictionaries*) and to address stemming errors due to conflicting and multiple rule-based word stemming to stem prefixation, suffixation and confixation words (*derivative dictionaries*). This proposed stemmer uses only dictionary entries of 5971 words compared to the existing Malay stemmers that uses complete dictionary with 22,481 words. As a result, the proposed stemmer minimizes dictionary lookup time to address the existing stemming errors [4-5][7].

```
1  Input: Accept the input text document
 2Remove HTML tags, special characters
 3Convert words from upper case to lower case
 4go to Step-1
 5
6Output: Root Words {stem1,stem2,stem3,....stemn}
 7
8Step-1: i = word1,word2,word3,...wordn
 9
10IF i = 0, go to Output
11IF i = wordn, identify word pattern and go toStep-
2
12
13Step-2: wordn = specific word pattern
14
15IFwordn = INFIXATION,
16check against lookup dictionary, stem wordn
17and go to Step-1
18
19IF wordn = CONFIXATION,
20Condition1:
21check against root word & derivative dictionaries
22IF wordn found, stem wordn and go to Step-1
23ELSEgo to next condition
24Condition2:
25stem wordn using confixation stemming rules
26go to Step-1
27
28IF wordn = PREFIXATION,
29Condition1:
30check against root word & derivative dictionaries
31IF wordn found, stem wordn and go to Step-1
32ELSE go to next condition
33Condition2:
34stem wordn using prefixation stemming rules
35go to Step-1
36
37IF wordn = SUFFIXATION,
38Condition1:
39check against root word & derivative dictionaries
40IF wordn found, stem wordn and go to Step-1
41ELSE go to next condition
42Condition2:
43stem wordn using suffixation stemming rules
44go to Step-1
```

Figure 1: Proposed Affixation Word Stemmer

V. Experimental Results And Discussion

In order to evaluate stemming accuracy of the proposed affixation word stemmer, 300 Malay online articles from Malaysiakini online portal have been used as testing dataset that consist of 58,563 word occurrences with 8,937 unique words. Out of total unique words, there are 3,369 affixation words [Dataset A] and 1,461 root words that are similar to affixation words [Dataset B]. There are two performance evaluations were conducted: word stemming evaluation of the proposed stemmer against affixation words using Dataset and word stemming evaluation of the proposed stemmer against root words that similar to affixation words using Dataset B. The first experiment indicates that the performance evaluation of proposed stemmer against affixation words achieved 98.34% stemming accuracy. There are three main factors that contributed to these stemming errors in the affixation word stemming against affixation words: combination of two words (*kesahketika*), misspelled words (*mengnigatkan*) and errors in character encoding (*noneketuanya*). None of these stemming errors falls under categories of Affixation Stemming Errors Type I, II, III, IV, V, VI, VII and VIII. These words are not affixation words but merely the issues of text documents formatting (character encoding in digital text documents). On the other hand, the second experiment indicates that the performance evaluation of affixation word stemming against root words that are similar to affixation words achieved 79.67% stemming accuracy. There are five main factors that contributed to high stemming errors in the affixation word stemming against root words that are similar to affixation words: names (*ali*), places (*jerman*), brands (*malaysiakini*), abbreviations (*miti*) and English words (*indian*). These stemming errors are not considered in this paper due to the research focus on word stemming against root words and not proper nouns such as names, places and brands. The results of these experiments are shown in Table 2.

Table 2
Experimental Results of the Proposed Affixation Word Stemmer

Experiment	Stemming Accuracy	Stemming Errors Samples
Proposed stemmer against affixation words using Data-set A	98.34%	1. Combination of two words: *agarmenghubungi, didedahseperti, keputusanperniagaan, kesahketika* 2. Misspelled words: *mengnigatkan, pera-kauan, perlatan, pertahanana, perum-paan, saebagai, sesetangah* 3. Errors in character encod-ing: *noneketuanya, nonetimbalan, non-ementeri*
Proposed stemmer against root words that similar to affixation words using Dataset B	79.67%	1. Names: *ali, amri, alvin, badawi, ban, chai, chan, deraman, dolah* 2. Places: *alor, iran, jeli, jerman, kaliman-tan, kedah, kelantan*

The experimental results indicate that the proposed stemmer successfully stems pre-fixation, suffixation, confixation and infixation words using 370 affixation stemming rules with dictionaries entries of 5971 words with promising stemming accuracy.

VI. Conclusion

This paper describes the proposed affixation word stemmer that addresses affixation stemming errors that occur in existing Malay word stemmers. This proposed word stemmer reduces eight different types of affixation stemming errors where the root causes of these stemming errors are due to the complexity of morphological rules in the Malay language. To address these stemming errors, the integration of an enhanced Rules Application Order stemming method with a Stemming Error Reducer were developed in this proposed word stemmer. Based on our experimental results, it can be concluded that the proposed word stemmer produces better stemming results to stem prefixation, suffixation, confixation and infixation words by addressing the root causes of the existing affixation stemming errors. Our future work will focus on elevating the proposed word stemmer to include a misspelled word checker and a dictionary that contains proper nouns (person's name and place) to further reduce stemming errors for further improvement on stemming accuracy.

Acknowledgment

The authors would like to thank the Editor-in-Chief and the anonymous reviewers of the manuscript for their valuable comments and suggestions. This research was funded by Universiti Teknologi Malaysia's Research University Grant PY/2014/02479.

References

[1] Abdullah, M. T., Ahmad, F., Mahmod, R., Sembok, T. M. T., "Rules frequency order stemmer for Malay language," *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 9, no. 2, pp. 433-438, 2009.

[2] Ahmad, F., Yusoff, M., Sembok, T. M., "Experiments with a Stemming Algorithm for Ma-lay Words," *Journal of the American Society for Information Science*, vol. 47, no. 12, pp.909-918, 2009.

[3] Alfred, R., Leong, L. C., On, C. K., Anthony, P.,"A Literature Review and Discussion of Malay Rule-Based Affix Elimination Algorithms," *The 8th International Conference on Knowledge Management in Organizations*, pp. 285-297, 2014.

[4] Darwis, S. A., Abdullah, R., Idris, N., "Exhaustive Affix Stripping And A Malay Word Register To Solve Stemming Errors And Ambiguity Problem In Malay Stem-mers," *Malaysian Journal of Computer Science*, 2012.

[5] Fadzli, S. A., Norsalehen, A. K., Syarilla, I. A., Hasni, H., Dhalila, M. S. S., "Simple Rules Malay Stemmer," *The International Conference on Informatics and Applications (ICIA2012)*, pp. 28-35, 2012.

[6] Hassan, A., " Morfologi," Vol. 13, 2006.

[7] Idris, N., Syed, S. M. F. D., "Stemming for Term Conflation in Malay Texts," *International Conference on Artificial Intelligence*, 2001.

[8] Kassim, M. N., Maarof, M. A., Zainal, A., "Enhanced Rules Application Order Approach to Stem Reduplication Words in Malay Texts," *Recent Advances on Soft Computing and Data Mining*, pp. 657-665, 2014.

[9] Leong, L. C., Basri, S., Alfred, R., "Enhancing Malay Stemming Algorithm with Background Knowledge," *PRICAI 2012 Trends in Artificial Intelligence*, pp. 753-758, 2012.

[10] Othman, A., *Pengakar Perkataan Melayu untuk Sistem Capaian Dokumen*. MSc Thesis. Universiti Kebangsaan Malaysia, Bangi, 1993.

[11] Ranaivo-Malancon, B., "Computational Analysis of Affixed Words in Malay Language," *Proceedings of the 8th International Symposium on Malay/Indonesian Linguistics, Penang*, Malaysia, 2004.

[12] Sankupellay, M., Valliappan, S., "Malay Language Stemmer," *Sunway Academic Journal*, vol. 3, pp. 147-153, 2006.

[13] Sembok, T. M. T., Yussoff, M., Ahmad, F.,"A Malay Stemming Algorithm for Information Retrieval," *Proceedings of the 4th*

International Conference and Exhibition on Multi-lingual Computing, Vol. 5, pp. 2-1, 1994.

[14] Sharum, M. Y., Abdullah, M. T., Sulaiman, M. N., Murad, M. A., Hamzah, Z. Z.,"MALIM - A new computational approach of Malay morphology," *International Symposium of Information Technology (IT-Sim)*, Vol. 2, pp. 837-843, 2010.

[15] Tai, S. Y., Ong, C. S., Abdullah, N. A.,"On Designing An Automated Malaysian Stemmer For The Malay Language," *Proceedings of the Fifth International Workshop on Infor-mation Retrieval With Asian Languages*, pp. 207-208, 2000.

[16] Yasukawa, M., Lim, H. T., Yokoo, H.,"Stemming Malay Text and Its Application in Au-tomatic Text Categorization," *IEICE transactions on information and systems*, vol.92, no. 12, pp. 2351-2359, 2009.