Sunway Academic Journal 3, 147–153 (2006)

MALAY-LANGUAGE STEMMER

MANGALAM SANKUPELLAY[a]
SUBBU VALLIAPPAN
University of Malaya

ABSTRACT

Stemming is the removal of affixes (prefixes and suffixes) in a word in order to generate its root word. The objectives of this research were to build a software stemmer that can stem any given Malay word, and to develop a standard stemming algorithm for the Malay language. The Malay language was chosen because a complete stemmer for this language is unavailable. Stemmers have a wide variety of applications, such as in information retrieval and machine translation. It is expected that when this system is fully developed, it will benefit users and customers tremendously.

Key words: Stemmer, stemming, Malay, Porter Algorithm.

INTRODUCTION

Stemming can be defined as the removal of affixes (prefixes and suffixes) in a word in order to generate its root word. A stemmer is a software system designed to stem words. Stemming may sound simple but it is not an easy or trouble-free process. Anything related to language manipulation is complex and the same applies to stemming.

Stemming has a wide range of applications in numerous fields. However, the principal use of stemmers is for information-retrieval purposes. One of the main problems involved in information retrieval is variations in word forms (Lennon et al., 1981). The most common types of variations are spelling errors, alternative spellings, multi-word constructions, transliteration, affixes, and abbreviations. One way to alleviate such problems is to use stemming. Information-retrieval systems use stemming to improve the matching algorithms.

Stemmers are also used in applications such as automated text processing, speech synthesis and recognition, machine translation, handwriting recognition, grammar checking, and sentence generation.

There are many stemmers for the English language but there is none that is dedicated to the Malay language. English-language stemmers are quite complete and thorough. Since less attention has been given to the Malay language, it is worthwhile to develop a stemmer for the national language of Malaysia.

This research is aimed at developing a complete stemmer in the Malay language, which is not available presently. The stemmer will be able to stem approximately 12,000 Malay words, including for example, "*kata ganda*" or dual words—a notable feature of this stemmer. The specific objectives of this research are:

E-mail: [a]mangalam@um.edu.my.

- To develop a complete stemmer which would stem any given Malay word.
- To come up with a standard algorithm for developing a stemming program in the Malay language.
- To provide a general stemming engine which can be used for other applications and systems.

STEMMING

Stemming can be categorized into four approaches: affix removal, successor variety, n-gram and table lookup. Affix-removal algorithms remove suffixes and/or prefixes from terms, leaving a stem. These algorithms sometimes also transform the resultant stem. There are two subgroups under this category: longest-match and simple-removal. Longest-match stemmers remove the longest possible string of characters from a term according to a set of rules. They can be iterative in nature, that is, suffixes are removed one after another; or they can use a longest-match algorithm, that is, if more than one suffixes match the end of a word, the longest one is selected. Simple-removal stemmers remove only plurals from a term. It is commonly accepted that suffix stripping makes stemming easier.

Successor-variety stemmers use the frequencies of letter sequences in a body of text as the basis of stemming. The n-gram method conflates terms based on the number of di-grams or n-grams that they share. Another approach to stemming is the use of table lookups. Terms and their corresponding stems can be stored in a table. Stemming is then done by looking up terms in the table.

ENGLISH-LANGUAGE STEMMERS

Nice Stemmer

Nice Stemmer is a stemmer for English words developed by Yang et al. (2004). It is regarded as one of the most complete stemmers that have been developed in the stemming field. Nice Stemmer is composed of four distinct stemmers—Simple Stemmer, Porter Stemmer, Inflectional Stemmer, and Combination Stemmer—each of which handles a specific part of the stemming process.

Text Stemmer

The Text Stemmer was developed by Fox and Fox (2002). It is a trivial English stemmer that is designed to handle character encoding of the target language as well as to deal with morphological and orthographical variations. Its main function is to provide an easy-to-use stemming system that mimics the way a person narrows down and refines his search for a particular root word.

Porter Stemmer

The Porter Stemmer was written by Martin Porter (1980). It uses the process of removing the commoner morphological and inflexional endings from words in English. It is also being used as part of a term normalization process, which is usually done when setting up information-retrieval systems.

PORTER STEMMING ALGORITHM

After a thorough analysis and comparison among currently available stemming algorithms, the Porter Stemming Algorithm (Porter, 1980) was chosen as the basis for our stemmer. The Porter algorithm involves a multi-step process that successively removes short suffixes, rather than removing in a single step the longest suffix. The algorithm is careful not to remove a suffix when the stem is too short, thus helping to improve the performance of the resultant stemmer.

The original Porter algorithm was intended for stemming in the English language. However, this algorithm can be implemented in the Malay language, too. The differences between these two languages necessitated some modifications in the stemming algorithm. For example, there are a few differences between word structures in Malay and English, such as:

- The combination of affixes attached to a word.
- The differences in syllables used to construct a word.
- The presence of infixes in Malay (which is absent in English).

Hence, the original Porter Stemming Algorithm was used only as a guide in the development process, and was not adapted totally. Our modified algorithm also aims to handle exceptional scenarios which are present in the Malay language. An overview of the algorithm is as follows:

```
Step1 (root_word);
Step2 (stem_suffix);
Step3 (stem_prefix);
Step4 (stem_infix);
Step5 (stem_others);
```

The detailed stemming algorithm is presented here:

Step 1
Check input word against the word dictionary
    If the word is found in the dictionary, then
        Output the word as root word
        Else,
        Go to Step 2

Step 2
Check input word for any prefix
    If the word has a prefix, then
        Remove the prefix and go to Step 1
    Else,
        Go to Step 3

Step 3
Check input word for any suffix
    If the word has a suffix, then
        Remove the suffix and go to Step 1
    Else,
        Go to Step 4

Step 4
Check input word for any infix
    If the word has a infix, then
        Remove the infix and go to Step 1
    Else,
        Go to Step 5

Step 5
Check stemmed word for its first letter
    If the first letter starts with ‘m’, then
        Replace the first letter with ‘p’ or ‘f’ and go to Step 1
    If the first letter starts with ‘n’, then
        Replace the first letter with ‘t’ and go to Step 1
    If the first letter starts with ‘y’, then
        Replace the first letter with ‘s’ and go to Step 1
    If the first letter starts with a vowel, then
        Add ‘k’ as the first letter and go to Step 1
    Else,
        Root word could not be found and the program terminates

RESULTS

Table 1 lists the supported affixes (prefixes and suffixes) of our stemming algorithm.

Table 1. Affixes Removable by the Stemming Algorithm

Prefix	'*ber*', '*per*', '*ter*', '*mem*', '*pem*', '*menge*', '*penge*', '*meng*', '*peng*', '*men*', '*pen*', '*me*', '*pe*', '*be*', '*ke*', '*se*', '*te*', '*di*'
Suffix	'*nya*', '*kan*', '*an*', '*i*', '*kah*', '*lah*', '*pun*', '*ita*', '*man*', '*wan*', '*wati*', '*ku*', '*mu*'
Prefix and suffix	'*ber…an*', '*per…an*', '*ter…kan*', '*mem…kan*', '*pem…an*', '*pen…an*', '*pe…an*', '*ke…an*', '*se…an*', '*te…kan*', '*di…kan*', '*ber…kan*', '*me…i*', '*men…i*', '*meng…i*', '*menge…kan*', '*penge…an*', '*peng…an*'
Two or more affixes	'*diper…*', '*…kannya*', '*memper…i*', '*berke…an*', '*men…inya*', '*di…kannya*'

Another advantage of this stemming algorithm is its capability to stem dual words or "*kata ganda*." The algorithm does not stop there; if there are any further affixes (both prefixes and suffixes) attached to this dual word, the stemming process continues thereafter to remove the attached affixes. Some results are shown in Table 2.

Table 2. Dual-Word Stemming

Dual Words ("Kata Ganda")	Sample	Stem
Words without a prefix (non-identical words)	saudara-mara	saudara
Words with a prefix but without a suffix	**ber**lari-lari	lari
Words with a prefix but without a suffix	**ter**tanya-tanya	tanya
Words with a prefix but without a suffix	**mem**beli-belah	beli
Words with a prefix but without a suffix	**men**deru-deru	deru
Words with a prefix but without a suffix	**meng**elak-elak	elak
Words with a prefix but without a suffix	**me**lihat-lihat	lihat
Words with a prefix but without a suffix	**se**akan-akan	akan
Words with a suffix but without a prefix	satu-satu**nya**	satu
Words with a prefix and with a suffix	**ke**anak-anak**an**	anak
Words with a prefix and with a suffix	**se**baik-baik**nya**	baik

This stemmer is also able to stem sentences rather than just words, giving added flexibility to the stemmer. Table 3 shows some test sentences stemmed.

Table 3: Test Sentences Stemmed

Test Sentence	Results
ahmad berjalan kaki ke perpustakaan	ahmad jalan kaki ke pustaka
pelajar-pelajar universiti akan menduduki peperiksaan pada minggu ini	ajar universiti akan duduk periksa pada minggu ini
separuh daripada hartanya didermakan kepada rumah anak-anak yatim	paruh daripada harta derma kepada rumah anak yatim
sofia memerlukan dua buah beg untuk dibawa ke perkhemahan	sofia perlu dua buah beg untuk bawa ke khemah
mengenali sesama sendiri adalah amat perlu untuk mewujudkan persefahaman mutlak	kenal sama sendiri adalah amat perlu untuk wujud faham mutlak

CONCLUSION

As the first stemmer to be developed specifically for the Malay language, this stemmer carries added significance for stemming in the Malay language. The stemmer is designed to handle all exceptions in the language. Its importance and benefits to the local community and industry are unquestionable. The scalability of this stemmer fully supports any future enhancements. In summary, this Malay-language stemmer is just the beginning of a linguistic advancement with a host of other applications waiting in the pipeline, such as:

1. **Online Dictionary**
   Currently the stemmer is able to provide the root word for a given word. This root word later can be incorporated into an online Malay dictionary to provide dictionary functionalities such as word definition, antonym, pronunciation, and example of usage.

2. **Language Translator**
   This feature is an expansion of the previous one. Once the definition of a word is found, it can be integrated with online dictionaries from other languages in order to operate as a language translator. Alternatively, other language dictionaries can be loaded into this system to perform the same function.

3. **Text-To-Speech System**
   Stemmers are widely used in the field of information retrieval but speech synthesis (text-to-speech) provides a new area of interest. This stemmer could be integrated with a few other applications to construct a text-to-speech synthesizer. The role of the stemmer is principally to solve the “e-pronunciation problem” in the text-to-speech synthesizer.

REFERENCES

Fox, B., & Fox, C. J. (2002). Efficient stemmer generation. *Inf. Process. Manage.*, 38*8*(4), 547-558.

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

Porter, M. 1980. An algorithm for suffix stripping. Retrieved 9 May 2006 from http://www.tartarus.org/martin/PorterStemmer/def.txt.

Yang, K., Song, D., Jeoung, W., & Tang, R. 2004. Nice Stemmer. Retrieved 24 March 2004 from http://ils.unc.edu/iris/irisnstem.htm.