Proceedings of The Fifth International Workshop on Information Retrieval with Asian Languages, Hong Kong, September 2000

On Designing an Automated Malaysian Stemmer for the Malay language

Sock Yin Tai, Cheng Soon Ong, Noor Aida Abdullah

Software Lab, MIMOS Berhad, Technology Park Malaysia,
57000 Kuala Lumpur, Malaysia.
Email: sytai@mimos.my

Abstract

Online and interactive information retrieval systems are likely to play an increasing role in the Malay Language community. To facilitate and automate the process of matching morphological term variants, a stemmer focusing on common affix removal algorithms is proposed as part of the design of an information retrieval system for the Malay Language. Stemming is a morphological process of normalizing word tokens down to their essential roots. The proposed stemmer strips prefixes and suffixes off the word. The experiment conducted with web sites selected from the World Wide Web has exhibited substantial improvements in the number of words indexed.

Keywords: information retrieval, stemming, world wide web, Malay Language

Introduction

According to a recent report by International Data Corporation, Internet users in the Asia-Pacific are continuing to show a clear preference for viewing the Web in their native language., especially when English is not their native language [1]. We in Malaysia are among the 215 million users of the Malay language within the South East Asian region [2]. Web pages in Malay are also showing exponential growth in numbers, similar to the trends exhibited by the size of the World Wide Web in general. Consequently, online and interactive information retrieval systems are likely to play an increasing role in the Malay Language community. In this paper, Malay refers to the official language of Malaysia, also known as Bahasa Malaysia or Bahasa Melayu. However, Bahasa Malaysia is very similar to Bahasa Indonesia, hence the arguments here apply to Bahasa Indonesia as well.

One technique for improving information retrieval performance [3] is to automate the process of matching morphological term variants. Thus, a stemmer focusing on common affix removal algorithms is therefore proposed as part of the design of an information retrieval system for the Malay Language.

Section 2 of this paper discusses the structure of Malay, and section 3 details the stemming process. A description of the proposed Malay stemmer is given in section 4, with the results obtained being discussed in section 5.

Conflation in the Malay Language

Linguistic variation is responsible for many possible alternative formulations of a single meaning. Each language has its own unique structure and grammar and the Malay language is no exception, with its own morphology or word formation process.

The Malay language has four main categories of affixes[8]. These are the prefix, suffix, "apitan" and "sisipan". Prefixes refer to the element added at the beginning of each root word. Suffixes refer to the element added at the end of each root word. "Apitan" refers to affixes which contain both prefixes and suffixes. Finally, "sisipan" refers to affixes which occur in the middle of the word. For example, "t*el*unjuk" (index finger) originates from the word "tunjuk" (to point). However, since "sisipan" usually changes the meaning of the word[8], our system will not stem them.

Stemming

Stemming is a morphological process of normalizing word tokens down to their essential roots, which is also known as conflation. Terms with a common stem will usually have similar meanings, for example: "buat" (make), "*di*buat" (made by) and "*mem*buat" (making).

In general, the stemming process consists of suffix and prefix removal. While there are many types of stemming algorithms available today, there are not any systems written specifically for the Malay Language

Stemmer Design

Many strategies for suffix stripping have been reported in the literature [5,6,7]. The nature of the task will vary considerably depending on whether a stem dictionary is being used, whether a suffix list is being used, and of course, on the purpose for which the suffix stripping is being done. The approach adopted here uses an affix list, with the aim of improving information retrieval performance. The affix-stripping program is given an explicit list of prefixes and suffixes. Along with each affix, the criterion under which it may be removed from a word to leave a valid stem is also included. It is to be noted that previous stemmer designs [5,6,7] have concentrated solely on suffix stripping. Due to the structure of the Malay Language, our stemmer performs prefix and suffix stripping.

The proposed system strips prefixes and suffixes off the word, once the respective criteria are met. This may occasionally result in a meaningless root word, but this would not affect the information retrieval process as both the indexed and query terms are stemmed in the same manner. Table 1 and 2 shows the prefixes and suffixes that have been stemmed by the system respectively.

Table 1: Prefixes stripped by the system

Bel	ber	di	diper	ke
Mem	memper	men	meng	menge
Meny	pel	pem	pen	peng
Penge	peny	se	ter	

Table 2: Suffixes stripped by the system

An	i	kan	man	Isme

Application notes

Certain heuristics were also included in the implementation of the stemmer program. These were obtained from the analysis of the behavior of the system when applied to real world data. It may be observed from table 1 that the prefixes are at least two characters in length and may be as long as six characters. Hence, words, which are shorter than five characters in length, are not stripped of their prefixes. Additionally, from table 2, suffixes are two characters to four characters in length, words that are shorter than five characters are will not be stripped of their suffixes.

An idiosyncrasy of the Malay language is that words beginning with “s” and “t”, in many cases, tend to lose the first character when prefixed by “me” and “pe” [8]. This has resulted in another heuristic in which “s” and “t” will stripped from the beginning of words.

Results and Discussion

To investigate the performance of the stemmer, three websites were selected to generate the list of words. These three sites gives rise to a of 93 different pages. The words in each page were extracted and stored, resulting in 22661 words in total. The three websites chosen are http:// dbp.gov.my/, http://www.lib.usm.my/press, http://cyberita.asia1.com.sg.

Out the 22661 words collected, 7994 unique words were obtained after elimination of duplicates. This list of words was then stemmed. The number of unique words obtained after stemming is 6191, contributing to a reduction of slightly more than 22% in the total number of unique words.

There are a number of issues in which we have not resolved. Firstly, as mentioned earlier, retrieval performance can be measured based on recall and precision. We have shown that stemming does improve recall, but have not concluded its effect on the precision of the query. Further work in refining the heuristics used for stemming may also result in better performance.

In conclusion, there is an increasing need for an information retrieval system for the Malay Language. An important component for such a system would be a conflation tool, such as our proposed stemmer, which allows matching of different words which are derived from the same root word. Hence, our system would be an important step towards an automated information retrieval system for the Malay language.

References

1. http://www8.techmall.com/techdocs/TS991104-4.html
2. http://www.sil.org/ethnologue/countries/Asia.html
3. Van Rijsbergen, C.J., “Information Retrieval”, 2nd ed., Butterworths, 1979
4. Frakes, W.B., R. Baeza-Yates, “Information Retrieval: Data Structures and Algorithms”, Prentice-Hall, 1992.
5. Porter, M.F., “An algorithm for suffix stripping”, Program, 14(3) :130-137, 1980.
6. Lovins, J.B., “Development of a Stemming Algorithm”, Mechanical Translation and computation Linguistics, 11 (1) March 1968 pp 23-31.
7. Dawson, J.L. “Suffix Removal and Word Conflation.”, ALLC Bulletin, Michaelmas 1974 p.33-46.
8. Nik Safiah, Farid, Hashim, “Tatabahasa Dewan”, 2nd ed. Dewan Bahasa dan Pustaka, Malaysia, 1989.