EVALUATION OF A SHALLOW STEMMING ALGORITHM FOR SINHALA

Viraj Welgama
Language Technology Research Laboratory,
University of School of Computing,
No. 35, Reid Avenue,
Colombo 07
www@ucsc.lk

Introduction

Stemming is an essential process in the field of Natural Language Processing (NLP). It is the process of reducing the inflected form of a word to its stem. Even though there are some advanced stemmers for languages such as English, the algorithms which they employ do not work well for highly inflectional languages such as Sinhala. The evaluation of a stemmer is complicated since it requires significant human effort. This poster describes a simple, shallow stemming algorithm and its evaluation against pre-defined lexical roots.

Motivation

Many NLP applications which use words as basic elements employ stemmers to extract the stems of words. This is a very efficient and lightweight approach compared to morphological parsing. On the other hand significant human effort and other language resources are needed to build complex linguistic resources such as morphological parsers. Computationally less resourced languages such as Sinhala lacks such linguistic resources and therefore call for alternate solutions such as efficient stemming algorithms to build these applications. This situation motivated us to carry out this work and evaluate how well a lightweight stemming algorithm can perform in comparison to a manually created lexicon.

Methodology

The Language Technology Research Lab (LTRL) at the University of Colombo School of Computing (UCSC) has developed a comprehensive Sinhala lexicon which covers over 83% unrestricted text obtained from online sources (Weerasinghe et al, 2009). We extracted 33,684 stem values and 1,325,273 corresponding word forms from this lexicon and created a *gold standard* for evaluating the proposed lightweight stemming algorithm which is defined as follows.

1. Obtain the list of words from the data set which needs to be stemmed
2. Sort the Wordlist alphabetically
3. Define the first word of the list as the stem value of itself
4. Check whether the second word starts with the first word and ends with one of suffixes in the suffixes list.
5. If yes, the stem of the second word is defined as the first word
6. Else the stem of the second word will be assigned as the second word itself
7. This process will be continued till end of the word list

Table 1.1 shows some of results generated from this algorithm.

Table 1.1: Some results of the lightweight stemming algorithm

Word	Stem	Word	Stem	Word	Stem
අ.පො.ස	අ.පො.ස	අංකල්	අංකල්	අංගනයක	අංගනය
අංක	අංක	අංකවලට	අංක	අංගප්‍රත්‍යංග	අංගප්‍රත්‍යංග
අංකද	අංක	අංකුර	අංකුර	අංගපුලාවක්	අංගපුලාවක්
අංකය	අංක	අංකුරවල	අංකුර	අංගම්පොර	අංගම්පොර
අංකයක්	අංක	අංග	අංග	අංගය	අංගය
අංකයද	අංක	අංගන	අංග	අංගයක්	අංගය
අංකයෙන්	අංක	අංගනය	අංගනය	අංගයකි	අංගය

Dataset

A list of distinct words extracted from the *UCSC 1M Words Sinhala News Editorial Corpus* was used to test the above stemming algorithm. 46,874 distinct Sinhala words were sorted alphabetically to identify their stem values using the above algorithm. A list of Sinhala suffixes was identified through the LTRL lexicon and then missing values were added in a trial-and-error basis. A total of 413 Sinhala suffixes were identified.

Experiments

The list of stem values generated using the proposed lightweight algorithm were compared with the defined gold standard. The efficiency of the algorithm was defined using the following equation.

$$Efficiency = \frac{No\ of\ Correctly\ Identified\ Stems}{Total\ Number\ of\ Words\ in\ the\ List} \times 100\%$$

26,272 stem values out of 46,874 words have been identified as real stems when compared with the gold standard. So the efficiency of the proposed algorithm is 56.04%.

Conclusion

- Over 50% of Sinhala words can be stemmed using a simple algorithm.
- The basic naïve algorithm used above can be enhanced using linguistic knowledge to improve the accuracy achieved.

Discussion

The gold Standard gives the real linguistic stem value which was defined manually. However many NLP applications which use stemming need not get the real stem value, but need to be able to cluster or identify the words which share the same stem. Therefore the usability of the proposed algorithm would be higher than the measured efficiency.

Reference

Weerasinghe, A. R, Herath, D. L, Welgama, V. "*Corpus-based Sinhala Lexicon*". 07th Asian Language Resource Workshop, ACL-IJCNLP 2009, Singapore – August 2009