Stemming Bangla

csohel@gmail.com,imran_cse_sust@yahoo.com, ayon_cse_sust@yahoo.com

Abstract

Stemming is the process of removing the affixes from inflected words to their stem, base or root form which is widely used in computational linguistics , information retrieval and natural language processing. Bengali is a highly inflected language. So for processing Bengali text in information retrieval stemming is needed. In this paper we introduce a corpus based and affix removal stemmer for Bangla language which is computationally inexpensive, simple and highly effective .It can reduce stemming errors (over-stemming and under-stemming).We have created a big database and a list of 300 suffixes and 56 prefixes. With the help of this database we are getting higher accuracy than the other stemmers. The result of proposed stemmer is encouraging which shows stemming can be performed with low error rates and will be effective in information retrieval system.

Keyword

Bengali, Stemmer, Corpus based Stemmer, Affix Removal, Over-stemming, Under-stemming, Stemming Algorithm, Information Retrieval (IR), Stemming.

Introduction

Stemming is a core natural language processing technique for efficient and effective Information Retrieval (Frakes 1992). It is used to improve the ability to match query and document vocabulary.With the rapid growth of Bengali data available in Internet, there is a practical need for developing a stemmer for processing Bengali language. This paper presents a corpus based and affix removal stemmer. There are several types of stemming algorithms. Affix removal algorithms are the most common. Affix removal stemming algorithms remove affixes (suffixes or prefixes) from words producing a root form called a stem. We have build a Bangla corpus for look up.We have also made a list of prefixes and suffixes.We have got 56 prefixes and 300 suffixes.In our stemming, first we look up the input word in the corpus ,then if we do not get the input word in the corpus we try to remove affixes from the input word .

Corpus-Based Technique

There are several approaches of stemming. In the corpus-based stemmers it matches every word with a word on a proper structured corpus, correspond each word to its stem. This procedure is effective but building a efficient corpus is a challenge.

Related work

Stemming has been extensively studied for English and many other European languages. In this section, we will give an overview of the related work on this problem.

The first ever published stemmer was written by Julie Beth Lovins in 1968. This paper was remarkable for its early date and had great influence on later work in this area. A later stemmer was written by Martin Porter and was published in the July 1980 issue of the journal Program. This stemmer was very widely used and became the de-facto standard algorithm used for English stemming. There are a number of other stemming algorithms for English such as Paice/Husk [7], Lovins Stemming [8], Dawson [9], and Krovetz [10].

Very limited works have been done in the past in the areas of stemming in Bengali language.There exists a light weight stemmer for Bengali[].There is a morphological stemmer based on stemming cluster technique has been developed for Bengali[].There is also an implementation of clustering based stemming algorithm by Barnan Das & Tanmoy Pal.[]

Bangla Corpus

In linguistics a corpus(plural corpora) is a large and structured set of texts.Corpus is considered as a basic resource for language analysis and research.So for Bangla language analysis and research work it is necessary to have a complete Bangla linguistic corpus. We have made a corpus for Bangla root word by a web crawler from ().We have also made a news corpus from "Kaler kanto" newspaper.

Bangla language Stemmer

Bengali is one of the most spoken languages (ranking sixth) in the world. There has been an exponential growth rate of Bengali web and digital text contents in the last two decades .So for the development of highly efficient IR systems a good Bangla stemmer is needed.

Stemming problem in Bangla

Bangla is highly inflected language where more than one inflection can be applied to the stem to form the word type. There are also a large number of compound words where two roots can join together to from a compound word. Those words may also have some morphological variants. Bangla nouns, adjectives and verbs are mostly inflected. There are nearly (10*5) forms for a certain verb word in Bengali as there is 10 tenses and 5 persons and a root verb changes its form according to tense and person. So verb is the most problematic area for stemming.

Prefix list

A **prefix** is an affix which is placed before the root of a word. We have got 56 prefixes from Bengali language grammar. These prefixes are mostly present with the inflected words. The prefix list is given below:

অ,অঘা,অজ,অপ,অব,অভি,অতি,অধি,অনু,অনা,আ,আন,আব,আম,আড়,ইতি,উৎ,উন,উপ,কম,কদ,কার,কি,কু,খাস,গর,দর,না,নি,নির,নিম,পরা,পাতি,পরি,প্র,প্রতি,ফি,ফুল,ব,বদ,বর,বে,বি,ভর,রাম,লা,স,সা,সু,সম,সমা,সাব,হর,হা,হাফ,হেড

Suffix list

A **suffix i**s an affix which is placed after the stem of a word. We have got 300 suffixes from Bengali language grammar. These suffixes are mostly present with the inflected words. Some example of suffixes from list is given below:

তাম,িতেছিল,িতেছিলেন,ছিলেন,িতেছিলে,ছিলে,তেছিলি,ছিলি,িতেছিলাম,ছিলাম,িআছিল,েছিল,িয়াছিলেন,েছিলেন,িয়াছিলে,িছিলে,িয়াছিলি,েছিলি,িয়াছিলাম,েছিলাম,িবে,িবেন,বেন,িবি,বি,িব,িও,ো,ায়,ান,াও,াইস,াস,াই, াইতেছে,াচ্ছে,াইতেছে,াছেন,াইতেছ

Stemming procedure:

```
1)If the input word get in corpus
           Do Pos Tag the input word       //(the input word is root word)
2)else
         Try to stem     //go to stemmer
         i)try to cut prefix
                   if prefix cut successful
                                     set prefix_flag=1;
        ii)try to cut suffix
                   if suffix cut successful
                                       set suffix_flag=1;
3)   if(prefix_flag==1&&suffix_flag==0)
         Return prefix & root word;
     if(prefix_flag==0&&suffix_flag==1)
          Return suffix & root word;
      if(prefix_flag==1&&suffix_flag==1)
           Return input word;
    if(prefix_flag==0&&suffix_flag==0)
          {
          Try to cut suffix
             If suffix cut successful
                        Return root word;
              Else
                 Return root word;
           }
```

Stemming error

There is a possibility to evaluate stemming by counting the numbers of two kinds of errors that occur during stemming, namely;

- **Under Stemming.**

  Understemming is an error where two separate inflected words should be stemmed to the same root, but are not

- **Over-Stemming**

  Overstemming is an error where two separate inflected words are stemmed to the same root, but should not have been

We have shown the example of errors

Over stemming:

তিনি=>তিন

জীবন=>জীব

মাজার=>মাজা

Under stemming:

আবছায়া=>আবছায়া but output should ছায়া

কাঙালিনি= >কাঙালিনি but output should কাঙাল

সঠিক=>সঠিক but output should ঠিক

Double valid stem:

আসবে আ+সবে/ADV আসবে আসব+ে/NN

প্রসঙ্গে প্র+সঙ্গে/PRP প্রসঙ্গে প্রসঙ্গ+ে/NN

অবশেষে অব+শেষে/ADV অবশেষে অবশেষ+ে/ADJ

সুরে সু+রে/INT সুরে সুর+ে/NN

ভরতে ভর+তে/NN ভরতে ভরত+ে/NN

ভরসার ভর+সার/ADJ ভরসার ভরসা+র S/NN

হাটের হা+টের/NN হাটের হাট+ের/ADJ

উপমহাদেশীয় উপ+মহাদেশীয়/ADJ উপমহাদেশীয় উপমহাদেশ+ীয় ADJ

বিমানে বি+মানে/NN বিমানে বিমান+ে/NN

প্রেমিক প্র+েমিক/ADJ প্রেমিক প্রেম+িক S/NN\ADJ

বলেই ব+লেই/NN বলেই বলে+ই/ADV

প্রভার প্র+ভার/ADJ প্রভার প্রভা+র S/NN

সুজিত সু+জিত/NN সুজিত সুজি+ত S/ADJ

সাবের সা+বের/ADJ সাবের সাব+ের/pfx

পরিশোধিত পরি+শোধিত/ADJ পরিশোধিত পরিশোধ+িত S/ADJ

আলোর আ+লোর/NN আলোর আলো+র S/NN

দরজায় দর+জায়/NN দরজায় দরজা+য়/NN

আলিক আ+লিক/NN আলিক আলি+ক S/NN

Discussion

This stemmer is run on the test set of 16000 words of the Bangla newspaper “Kaler kanto”. The result is encouraging . Using this stemmer in information retrieval applications search as a search engine will be more effective . We have to determine the tradeoff between under-stemming and over-stemming. Bengali has a small number of prefixes. In our stemmer we also handle prefixes as well.

Conclusion

Bengali is a well-known and renowned language . We can achieve to the goal of information searching by Bengali language and motivate general people toward the use of Information Technology. With the worldwide proliferation of the Internet, increasing amounts of information are becoming available online in languages that have not received much attention from the IR/NLP community and for which language resources are scarce. For this available information to be useful, it has to be indexed and made searchable by an IR system. Stemming is one of the basic steps in the indexing process. In this paper, we have proposed a stemmer that is corpus-based.