Abstract

Stemming is the prerequisite step in Text Mining, Spelling Checker applications as well as a basic requirement for Natural Language Processing (NLP) tasks. Also it is very important in most of the Information Retrieval (IR) systems. This paper describes an affix stripping technique for finding out the stems from context free text in Nepali Language using lexical lookup based and rule based approach. It starts by introducing different types of lexicon, the basic unit of Nepali stemmer and few rules to identify the word in the lexicon. These rules and lexicons are applied in the design and implementation of an extensible architecture of a stemmer system for Nepali text. Finally designed stemmer performance is evaluated over different domains of 1,800 words. These domains include news on Economics, Health & Political in Nepali language, which are based on Devanagari Script. The overall accuracy of the designed system is 90.48%. Due to the absence of extensive linguistic resources, this technique shows improvement in the performance over simple rule based system.

References
- http://www. mpp. org. np
- http://www. bhashasanchar. org
- http://tdil. mit. gov. in

**Index Terms**

Computer Science  
Artificial Intelligence

**Keywords**

Stemmer; Lexicon; NLP; Text Mining; Spelling Checker; IR