Hindi Stemmer @FIRE-2013

Anubha Jain[1] and Sujoy Das[2]

[1]Research scholar, Department of Computer Applications, MANIT, Bhopal
[2]Associate Professor, Department of Computer Applications, MANIT, Bhopal

(jain05anubha@gmail.com, sujdas@gmail.com)

Abstract. This paper describes a language independent approach for extracting Hindi morpheme from a given list of Hindi words of morpheme extraction task (MET) at FIRE 2013. In this approach list of Hindi word is submitted to the system and it generates stemmed Hindi root word from it. The proposed approach has shown an improvement of 3.40% over baseline result. The approach is applied on Hindi language but it may also be used for any other language due to its language independent nature.

Keywords. Stemmer, Prefixes, Suffixes, Morpheme.

1 Introduction

In many of the domains where natural language is processed, depending upon the task, obtaining root or stem word from its inflected form may prove to be one of the most important preprocessing tasks. The process may be used to control number of index term that may be present at the time of indexing or may be useful at the time of query formulation. Controlled vocabulary at the time of indexing may improve retrieval performance [1]. Stemmer and lemmatizer are used to extract the lexical and grammatical morphemes of the original words [4].Morpheme extraction are a technique that is used for extracting morphemes from the inflected form of words. The system is referred as stemmer which is used for getting the stem, base or root form of any word. The morpheme extraction results in two classes of words i) Stem and ii) Affixes (Prefix, suffix etc.). Stem words are root/base form whereas affixes are used to modify the meaning or grammatical function of root word [2]. Hindi has large number of affixes rather than English [2]. In Hindi **सदस्य**, **बल्लेबाज**, **दृश्य**, **आश्रम** are root words and some of the affixes are **यों**, **ओं**, **अन**, **अनीय**, **अति**, **अन्** etc.

Stemmer can be rule based or statistical. In rule based stemmer predefined rules are used to design stemmer whereas statistical information of the corpus may be used to obtain root or base form [3][5]. Number of stemmers like Porters, Krovetz and Lovins [6][7], with their own pros and cons are used in English language but there is a need to develop a mature Hindi language stemmer.

2 Methodology

In this paper an attempt is made to extract root/stem word from list of words of Hindi language. The approach is language independent as it does not use additional list of root words, stop words or affixes. The approach is based on following assumption:

a) List of Hindi word should contain root word/base form as well as its morphological variants.

b) At the time of sorting, root word/ base form shall bubble up and is placed before all its morphological variants or inflected forms. **सदस्यों,सदस्य,सदस्यता**, **सदस्यीय**, shall be sorted as **सदस्य**, **सदस्यता**, **सदस्यीय**, **सदस्यों**.

The steps of algorithm are as follows:

1. **Stop words removal**: Words in Hindi word list with length less than 3 are considered as stop words and are removed from the list at time of preprocessing.
2. **Sorting single column word file**: Bubble sorting algorithm is applied for sorting words present in single column word file with N number of words.
3. **Stemming performed on sorted list**:
   i. Each word is compared with next 10 words assuming that maximum of 10 morphological variants is present in list.
   ii. If word is present as substring in next word then the word is broken as substring + remaining characters of word.
   iii. Substring is treated as root/base form and remaining characters are treated as affix.

3 Experiment and Result

The list of 407560 Hindi words is supplied on MET task at FIRE 2013. Our proposed approach could obtain 126937 base forms from this list. The result is shown in Table 1.

Word from Hindi word list	Base form
बालिकाएं	बालिका
सरकारिया	सरकार
बुराईयों	बुराई
मनुष्यता	मनुष्य
अदालती	अदालत
अदाकारा	अदाकार
सम्वेदनाओं	सम्वेदना

Table 1. Inflected Hindi words and its root/base form

Proposed approach is able to remove only suffixes in few cases. The result of suffix removal is shown in Table 2.

Hindi word	Base form	Suffix
अदृश्यता	अदृश्य	ता
अनादरणीय	अनादर	णीय

Table 2. Suffix removal

Some of the morphological variants could not be converted to its base form Table 3.

Hindi Word	After Stemming
खिलाड़ियों	खिलाड़ियों
प्रत्याशिओं	प्रत्याशिओं

Table 3. Unstemmed words

In some of the cases due to the presence of prefixes in Hindi word list, our system is able to extract prefixes also (Table 4).

Hindi Word	After Stemming
अंडरएक्टिव	अंडर
अंडरएज	अंडर
अंतरराष्ट्रीय	अंतर

Table 4. Prefixes from Hindi word list

The retrieval performance is shown in Table 5 and an improvement of 3.40% is reported over baseline. The baseline is 0.2821 whereas MAP with our approach is 0.2917.

Institute	Language	Baseline	MAP Obtained	% improvement
MANIT	Hindi	0.2821	0.2917	3.40%

Table 5. MAP and % improvement

4 Conclusion

This is our first attempt at MET task of FIRE 2013. A language independent approach is proposed. The proposed approach does not use an additional list of stop words or affixes and is able to improve the Mean Average Precision by 3.40% improvement over baseline. It is observed that in few of the cases some of the words are having more than 10 morphological variants in the list due to that our stemmer could not find all the variants. The approach is tested on only Hindi language and in future we shall try to test it on other languages.

References

1. Baeza-Yates, Ricardo, and Berthier Ribeiro-Neto. *Modern information retrieval*. Vol. 463. New York: ACM press, 1999.

2. www.cse.msu.edu/~cse842/Classnotes/Lecture2-Morphology.pdf

3. Ganguly, Debasis, Johannes Leveling, and Gareth JF Jones. "DCU@ FIRE-2012: Rule-based Stemmers for Bengali and Hindi."

4. Sahoo, B., M. Swain, and D. K. Sahoo. "IIITBh FIRE 2012 Submission: MET Track Odia."

5. Yadav, Avinash, Robins Yadav, and Sukomal Pal. "ISM@ FIRE-2012 Adhoc Retrieval Task and Morpheme Extraction Task."

6. Porter, M.F.: An algorithm for suffix stripping. Program 14(3) (1980) 130–137

7. Lovins, Julie Beth (1968). "Development of a Stemming Algorithm". *Mechanical Translation and Computational Linguistics* 11: 22–31.