Volume 4, Issue 1, January 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Hindi Rule Based Stemmer for Nouns

Vishal Gupta

UIET, Panjab University Chandigarh,

India

Abstract— Stemming is a technique for converting derived terms into corresponding root or stem words. It is not necessary that stem terms should be similar to root of that term. A typical English stemmer can convert different variants catty, cats and catlike into root term cat. Another example of this stemmer is to convert stemmed, stemmer and stemming into base form stem. A stemmer can perform operation of converting morphologically identical words to root word without performing morphological analysis of that term. A particular kind of very simple algorithm for performing stemming includes elimination of endings of terms by using suffix list which are used frequently. On the other hand a complex stemmer can apply knowledge related to morphology of that term for obtaining stem terms from derived words. Regarding Indian languages, very less work has been discussed for performing stemming of terms. But much of work has done for stemming in English and other European languages. Hindi is national language in India. Not much linguistic resources are available for Hindi as research is still at early stage for Hindi. This paper discusses Hindi stemmer for nouns. This Hindi stemmer applies suffix stripping rule based approach for performing stemming of nouns.

Keywords— Hindi stemmer, Hindi stemming, suffix stripping, rule based stemming, nouns stemmer

I. INTRODUCTION

Objective of any stemmer [1] is to convert terms into their basic root forms which are not in their base forms and are absent in dictionary. After performing stemming, If stemmed term is found in dictionary of that language, then we can say that the term is genuine. On the other hand if the stemmed term is not found in dictionary of that language then that term can be any named entity or invalid term. It is technique of converting derived terms into corresponding root or stem words. It is not necessary that stem terms should be similar to root of that term. A typical English stemmer can convert different variants catty, cats and catlike into root term cat. Another example of this stemmer is to convert stemmed, stemmer and stemming into base form stem. A stemmer can perform operation of converting morphologically identical words to root word without performing morphological analysis of that term.

A particular kind of very simple algorithm for performing stemming includes elimination of endings of terms by using suffix list which are used frequently. On the other hand a complex stemmer can apply knowledge related to morphology of that term for obtaining stem terms from derived words. Regarding Indian languages, very less work has been discussed for performing stemming of terms. But much of work has done for stemming in English and other European languages. Hindi is national language in India. Not much linguistic resources are available for Hindi as research is still at early stage for Hindi. This paper discusses Hindi stemmer for nouns.

II. RELATED WORK OF STEMMING FOR INDIAN LANGUAGES

Ramanathan and Rao (2003) [2] developed a stemming algorithm for Hindi which was lightweight stemming approach which applied a set of rules of suffix stripping based on largest match. This Hindi stemmer has applied list of sixty five suffixes which were created manually. It was tested on text documents taken from different fields like entertainment, politics, business, health and sports. The number of unique terms were thirty five thousand nine hundred seventy seven. This algorithm has discussed errors related to over stemming of around fourteen percent and under stemming of around five percent. It has not reported any recall or precision for performing stemming. Islam et al. (2007) [3] discussed a stemmer for Bengali which is light weight in nature. This algorithm removes suffixes by applying list of predefined suffixes for Bengali. It used seventy two verb suffixes, twenty two suffixes for nouns and eight for Bengali adjectives. Majumder et al. (2007) [4] proposed a stemmer called YASS (Yet Another Suffix Stripper) which used a statistical based clustering technique, on the basis of measure of string distance. It does not require any language oriented knowledge. This approach suggested that with stemmers, we can attain better recall for information retrieval methods in Indian languages i.e. Bengali language. Dasgupta et al. (2006) [5] discussed morphological parsing which is unsupervised for Bengali. It is job of splitting terms to stems, prefixes and suffixes in absence of language oriented rules related to morpho-phonological and morphotactics. There are two phases of this approach: i) presence of suffixes, prefixes, and roots words from lengthy corpus which is unannotated. ii) splitting terms on the basis of morphemes. This system attained F-measure of eighty three percent. Pandey et al. (2008) [6] discussed a Hindi stemmer which is an unsupervised in nature. For performing training, terms belonging to documents in Hindi from corpus called EMILLE were taken. After this, probabilities related to stem and suffix were

determined. Split probability was obtained from them. Steps related to post processing were performed for refining suffixes learned. Accuracy of this approach is around ninety percent with F-measure of around ninety five percent. Majgaonker et al. (2010) [7] discussed a stemmer for Marathi using unsupervised technique. For creation of suffix rules, It has applied 03 techniques: i) statistical stripping ii) stemming based on rule based approach and iii) stripping of suffixes. This rule oriented stemmer applies different manual rules of suffix stripping but on the other hand unsupervised technique automatically learns different types of suffixes from different terms taken from Marathi. Efficiency of this technique is around eighty two percent for stripping of suffixes based on statistical approach . Suba et al. (2011) [8] discussed Gujarati stemmer based on 02 types of approaches: i) lightweight approach of stemming on basis of hybrid technique and ii) heavyweight approach of stemmer on basis of rule oriented technique. Efficiency of light weight stemmer is around ninety one percent. Efficiency of heavy weight stemmer is around seventy one percent. Gupta & Lehal (2011) discussed rule based stemmer of nouns and proper names for Punjabi. In this stemmer effort was devoted for creating stem of words in Punjabi and then this stem is searched in Punjabi morph related to nouns and names. Eighteen suffixes were developed for nouns and names in Punjabi and overall efficiency of this stemmer is around eighty seven percent tested on corpus of Punjabi news [9].

III. HINDI NOUNS STEMMING

We can call nouns as names of place, person and concept. Nouns are very important in many applications of natural language processing and text mining. After thorough analysis of Hindi text taken from different news papers from Hindi, 16 suffixes of Hindi nouns have been obtained as given in TABLE I:

TABLE I
SUFFIXES OF HINDI NOUNS

Serial No.	Hindi Noun suffix	Suffix Example	Hindi Root word
1	यां	दवाईयां "Medicines"	दवाई "Medicine"
2	े	लडके "Boys"	लडका "Boy"
3	ों (with two variants)	फलों "Flowers" (1st Variant) कपड़ों "Clothes" (2nd Variant)	फल "Flower" (1st Variant) कपड़ा "Cloth" (2nd Variant)
4	यों	दवाईयों "Medicines"	दवाई "Medicine"
5	ऐं	कक्षाऐं "Classes"	कक्षा "Class"
6	ओं	कक्षाओं "Classes"	कक्षा "Class"
7	ि यों	लडकियों "Girls"	लडकी "Girl"
8	ि यां	लडकियां "Girls"	लडकी "Girl"
9	ेंं	पतंगें "Kites"	पतंग "Kite"
10	ए	कछुए "tortoises"	कछुआ "Tortoise"
11	एँ	माताएँ "Mothers"	माता "Mother"
12	ं	चिड़ियां "Sparrows"	चिड़िया "Sparrow"
13	ि यां	बिल्लियाँ "Cats"	बिल्ली "Cat"
14	ि यों	बिल्लियों "Cats"	बिल्ली "Cat"
15	ि ओ	लडकिओ "Girls"	लडकी "Girl"
16	ि ओं	लडकिओं "Girls"	लडकी "Girl"

Procedure of Stemming for Hindi Nouns

Step I: Enter Input Text in Hindi

Step II: Segment this text into words and search each Hindi word in Hindi Word-Net [10].
If that word is found in Hindi Word-Net and is tagged as noun then word is itself in root form.
Else If word is not found in Hindi Word-Net then go to Step III for performing its stemming

Step III: If Suffix of that word matches with any of suffixes: यां or ों or यों or ऐं or ओं or ेंं or एँ or ं
Then eliminate this suffix from end of that word and this stemmed word is again searched in Hindi Word-Net for noun possibility. If word is tagged as Noun then that word is returned as result. Otherwise go to Step-IV.

Step IV: If Suffix of that word matches with any of suffixes: े or ो

Then eliminate this suffix from end of that word and add ा at end of stemmed word.

Search this resulting word in Hindi Word-Net for noun possibility. If word is tagged as Noun Then that word is returned as result. Otherwise go to Step-V.

Step V: If Suffix of that word matches with any of suffixes: ियों or ियां or ि यों or ि यां or िओ or िओं

Then eliminate this suffix from end of that word and add ी at end of stemmed word.

Search this resulting word in Hindi Word-Net for noun possibility. If word is tagged as Noun Then that word is returned as result. Otherwise go to Step VI

Step VI: If Suffix of that word matches with ए

Then eliminate this suffix from end of that word and add आ at end of stemmed word.

Search this resulting word in Hindi Word-Net for noun possibility. If word is tagged as Noun Then that word is returned as result. Otherwise go to Step VII

Step VII: The word is not Hindi Noun.

Procedure Input: दवाईयां "Medicines", चिड़ियां "Sparrows", लडकियों "Girls", बिल्लियाँ "Cats"

Procedure Output: दवाई "Medicine", चिड़िया"Sparrow", लडकी "Girl", बिल्ली "Cat"

IV. IMPLEMENTATION AND RESULTS

We have implemented this procedure for stemming of Hindi nouns in VB.NET at front end and access at back end. This stemmer has been tested on 100 news documents of Hindi taken from popular Hindi news papers such as http://navbharattimes.indiatimes.com, www.amarujala.com, and www.punjabkesari.in. The accuracy of this stemmer is 83.65%.

Accuracy of Hindi Noun Stemmer= Number of Correctly stemmed Hindi noun Terms/ Total No. of Stemmed Words Errors of 16.35% are due to absence of some noun suffixes in this Hindi stemmer and also because of absence of some noun terms in Hindi Word-Net, violation rules of Hindi noun stemming or due to syntax errors in input text.

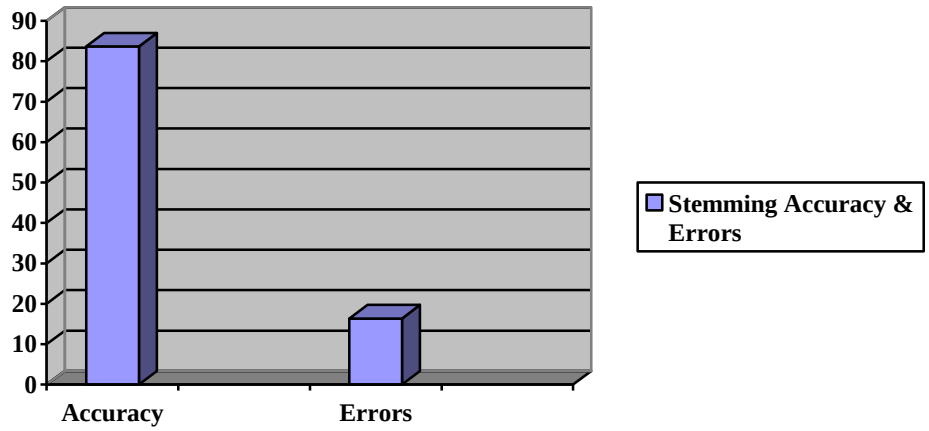

Fig. 1 Accuracy and errors of Hindi noun stemming

V. CONCLUSIONS

This stemmer for Hindi nouns is applying rule based suffix stripping technique. After analyzing news articles from popular Hindi news papers, 16 noun suffixes have been generated and corresponding stemming rules were developed for each suffix. Hindi noun stemmer is a basic linguistic resource which is essential for many applications in filed of text mining and natural language processing like: Text Summarization, Machine Translation, Keywords Extraction, Topic Tracking, Documents Clustering and Classification etc. The accuracy of this stemmer can be improved by adding more suffixes of Hindi nouns and more stemming rules.

REFERENCES

[1] V. Gupta and G.S. Lehal, "A Survey of Common Stemming Techniques and Existing Stemmers for Indian Languages," *Journal of Emerging Technologies in Web Intelligence*, vol.5, pp.157-161, 2013.

[2] A. Ramanathan and D. D. Rao, "A Lightweight Stemmer for Hindi," *Workshop on Computational Linguistics for South-Asian Languages*, EACL, 2003.

[3] M. Z. Islam, M. N. Uddin and M. Khan, "A Light Weight Stemmer for Bengali and its Use in Spelling Checker," *International Conference on Digital Communication and Computer Applications (DCCA07)*, Jordan, 2007.

[4] P. Majumder, M. Mitra, S. K. Parui, G. Kole, P. Mitra, and K. Datta, "YASS: Yet Another Suffix Stripper," *Association for Computing Machinery Transactions on Information Systems*, vol.25, pp. 18-38, 2007.

[5] S. Dasgupta and V. Ng, "Unsupervised Morphological Parsing of Bengali," *Language Resources and Evaluation*, vol. 4, pp. 311-330, 2006.

[6] A. K. Pandey and T. J. Siddiqui, "An Unsupervised Hindi Stemmer with Heuristic Improvements," *Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data*, vol. 303, pp. 99-105, 2008.

[7] M. M. Majgaonker and T. J Siddiqui, "Discovering Suffixes: A Case Study for Marathi Language,", *International Journal on Computer Science and Engineering*, vol. 02, pp. 2716-2720, 2010.

[8] K. Suba, D. Jiandani and P. Bhattacharyya, "Hybrid Inflectional Stemmer and Rule-based Derivational Stemmer for Gujarati," *Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), IJCNLP 2011*, Chiang Mai, Thailand, pp. 1-8, 2011.

[9] V. Gupta and G. S. Lehal, "Punjabi Language Stemmer for Nouns and Proper Names," *Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), IJCNLP 2011*, Chiang Mai, Thailand, pp. 35–39, 2011.

[10] http://www.cfilt.iitb.ac.in/wordnet/webhwn

